Balancing Stakeholder Needs in Adolescent mHealth Personalization Using Multi-Objective Reinforcement Learning

Shirley Li

TU Delft

Balancing Stakeholder Needs in Adolescent mHealth Personalization using Multi-Objective Reinforcement Learning

by

Shirley Li

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday, July 15th, 2026 at 11:00 AM.

Student number:	5026555	
Project duration:	November 10, 2025 – July 15, 2026	
Thesis committee:	Prof. dr. ir. W.P. Brinkman,	TU Delft, supervisor
	Drs. E.C.S. de Groot,	TU Delft, daily co-supervisor
	Dr. ir. U.K. Gadiraju,	TU Delft, committee member

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Acknowledgements

As I am writing this, it feels somewhat strange yet exciting to realize that this thesis marks the end of my student chapter at the TU Delft. Over the past seven years, I have learned a tremendous amount, not only in terms of technical knowledge and academic skills in computer science, but also through developing soft skills such as collaboration and organization. I also learned where my passions lie and how I want to apply my knowledge to societal problems. This thesis presented a fitting final project where these skills and interests came together.

There are several people whom I would like to acknowledge and thank for supporting me throughout this process. First and foremost, I would like to express gratitude to my supervisors, Esra and Willem-Paul. You helped me in times when I was unsure what to do; even though you were not the oracle that had the answer to my problems, you provided me with guidance and confidence to make my own decisions, which was much more valuable. Your feedback was also much appreciated, and you challenged me to see the bigger picture when I needed it.

I would also like to thank my friends and partner for providing a listening ear when things were not going as smoothly as planned, and for being the distraction needed at times. Of course, when there was good news I could also come to you to share my excitement, and you always celebrated with me. Thank you for the fun times together, and I am sure these moments will continue much further than our student time.

Next, I want to thank my family for supporting me throughout not only the process of writing this thesis, but also throughout all my student years. You were always there for me, provided me with care when needed, and reminded me that at some point it is good enough.

Lastly, I would also like to acknowledge the broader project context in which this thesis was conducted. This research builds upon the Grow It! application, which was developed as part of the flagship project PROTECt ME, an interdisciplinary collaboration between Erasmus Medical Centre Rotterdam, Erasmus University Rotterdam, and Delft University of Technology, aiming to support the mental health of adolescents. The efforts and contributions of everyone involved in this project created a valuable foundation for this thesis.

Shirley Li
Delft, July 2026

Abstract

To support adolescents' well-being, mobile health (mHealth) applications aimed at promoting habits that improve well-being and prevent mental health problems are being increasingly developed. To increase uptake and sustained usage, advancements in personalization for these applications have been made. However, personalization is not straightforward, as it should consider the diverse and sometimes conflicting needs of stakeholders: adolescents may prioritize enjoyment and desire low time and effort, while clinical experts emphasize adherence and usefulness. Existing reinforcement learning (RL) approaches either only optimize for one of such outcomes, or collapse several objectives into a single scalar number. As a result, these methods may fail to adequately capture and address all stakeholder needs, especially when objectives conflict. This thesis explores the use of multi-objective reinforcement learning (MORL) for personalization in an mHealth application for adolescents that is aimed at promoting four different coping categories. Through literature, we identified relevant needs and concerns from adolescents, clinical experts, parents, and society, and translated these into the design of our personalization algorithm through model components and algorithmic choices. We implemented a model-based MORL algorithm based on decomposition (MORL/D) with policy iteration, using data from 317 participants from an existing study. This produces a set of Pareto optimal policies, each representing a different trade-off between the objectives. We further proposed and evaluated three metapolicies for selecting from this Pareto set: user choice, expert priority, and a combination of both. Simulation-based evaluations revealed that conflicting objectives exist and that no single-objective policy is able to improve all outcomes simultaneously, revealing inherent trade-offs across and within stakeholders. In contrast, the user choice metapolicy that provides users with three options from the Pareto set achieved consistent improvements across all outcomes compared to a non-personalized baseline, while addressing adolescents' desire for autonomy and control by allowing them to select from a set of good options. This approach enhances user experience, engagement, and diversity of the practiced coping strategies and thus may better equip adolescents with a broad set of coping skills. These results support the use of MORL and demonstrate it can be a promising approach for mHealth personalization when balancing diverse and potentially conflicting stakeholder needs.

Contents

1 Introduction

Globally, one in seven adolescents battles with mental health conditions, such as depression, anxiety, or behavioral disorders [120]. Experiencing mental disorders during this age period can have major consequences for later stages in life, such as increased health problems [64, 97] and higher risk of unemployment [25]. Therefore, adolescents should be supported in developing habits that can help them with their well-being and prevent mental health problems. A valuable strategy to achieve this is by learning adaptive coping skills [71], which can help adolescents manage negative events. Studies by Frydenberg et al. [47] and Esmaeilimotlagh et al. [40] demonstrated that learning coping skills can positively affect adolescents' mental health in school and university, respectively. This aligns with the World Health Organization's recommendation to support adolescents in building resilience and developing adaptive coping strategies [120]. In recent years, mobile apps have become a popular way ta popular way to support adolescents' mental health and well-being [38, 72, 27]. One example is the Grow It! app [35], a serious gaming application for adolescents that combines emotion monitoring with daily cognitive behavioral therapy-based challenges to strengthen coping skills. While the app shows promise in improving adolescents' well-being, low user engagement remains a challenge. This reflects a broader issue in mHealth applications, where real-world uptake and sustained usage are low [76, 89, 116]. Personalization has been identified as a promising strategy to improve engagement, as it allows the application to match the needs and circumstances of individual users [117, 10, 127]. Yet, personalization is not straightforward, as there may be multiple needs that should be considered. For example, adolescents may want an application that is enjoyable, useful and not too time-consuming, whereas clinicians may put greater emphasis on effectiveness and safety. As a result, personalization must balance these multiple needs and objectives. Building on these insights, this study investigates how the coping skill-enhancing challenges in the Grow It! app can be personalized while accounting for diverse needs and concerns.

One method to implement personalization in mental health apps is by the use of reinforcement learning (RL) [111], which Weimann et al. [118] showed to be effective in various digital healthcare contexts. In RL, the system observes a user's current state (e.g., mood, tiredness, or motivation) and selects an action, such as recommending a coping challenge. After the action is taken, the system receives feedback in the form of a reward signal that reflects how desirable the outcome was. The goal of the algorithm is to learn a decision-making strategy (policy) that maximizes the expected cumulative reward over time. This reward typically measures a proximal outcome, such as an increase in positive affect, which serves as a measurable proxy for the distal outcome [88], such as long-term improvements in mental well-being. By optimizing for this reward, the RL algorithm can learn a policy (i.e., a mapping from states to actions) that makes personalized recommendations tailored to users' needs.

However, optimizing for a single reward may not adequately capture the full personalization problem, as there may be multiple, possibly conflicting objectives. For example, adolescents might want coping challenges that are useful, but do not require too much effort. Or they may want a challenge to be enjoyable, but not too time consuming. Additionally, different stakeholders in such might have different interests as well. For example, a study by Terlouw et al. [114] found that children wanted a game that made them feel connected with peers, while health care professionals were more focused on developing social skills. Moreover, what is considered important by an expert often does not align with

what users perceive as useful [3]. These varying interests illustrate that a single-objective approach may be insufficient, since optimizing only for the clinical outcome may overlook the user-centered factors that sustain long-term engagement. This motivates the need for methods that can represent and optimize multiple objectives.

A common approach to deal with multiple objectives in RL, is to combine the several reward signals into a single objective using a linear scalarization function [59]. This effectively transforms the multi-objective problem into a single-objective problem and makes it possible to apply standard RL algorithms. However, defining this scalarization function requires a-priori knowledge about the preferences of the several objectives. For instance, if we want to address effectiveness and enjoyment, we would need to know how much weight to give to each of these objectives. In practice, it is often difficult to define these preferences, and they can change over time [59]. For example, an adolescent might prioritize enjoyment on a day when they feel demotivated, but prefer shorter, more effective challenges on a busy or tiring day. Consequently, a ”one-size-fits-all” scalarization may produce suboptimal results as it cannot adapt to these fluctuating preferences.

Multi-Objective Reinforcement Learning (MORL) [59] addresses this problem by accounting for multiple objectives without a predefined trade-off. Instead of searching for a single optimal policy, it finds a set of optimal policies that represent the best possible trade-offs, resulting in multiple optimal actions per state. Relating this to the coping skill-enhancing challenges, we can utilize MORL to find a set of challenges that are optimal in different ways. For example, it can propose a challenge that is less time-consuming but perhaps less effective, and another challenge that is more effective but less enjoyable. The final selection can then be made through a higher-level decision mechanism, such as allowing adolescents to choose between a set of challenges, or in collaboration with healthcare experts. By removing the trade-off decision from the personalization algorithm, we allow for more flexibility in the final selection and hope to better align with the fluctuating and diverse needs of the involved stakeholders.

MORL has been applied in various problem domains where multiple aspects need to be considered. For example, in robotics, where speed and energy efficiency need to be balanced [122], and in operations of a water resource system, where we want to maximize the energy production of the hydropower plant, while minizing the risk of flooding [20]. Some applications in healthcare exist as well: for example, Lizotte et al. [79] developed an MORL algorithm for finding the best treatment plan for schizophrenia by balancing symptom reduction against the severity of side effects, and Jalalimanesh et al. [62] proposed an MORL algorithm for calculating the radiotherapy dose while minimising treatment time, total radiation dose and healthy tissue damage simultaneously. These adaptations of MORL in clinical domains demonstrate their ability to address multiple, often conflicting demands in these settings. This suggests it can be applied in mHealth personalization as well, to account for the diverse needs and concerns. Therefore, this research aims to investigate the potential of MORL for the personalization of coping skill-enhancing challenges within a mental health app for adolescents.

1.1. Research Question

Understanding and addressing the diverse needs and concerns of the involved stakeholders is necessary for encouraging sustained usage of mental health apps [31]. Current Reinforcement Learning solutions often simplify these varying interests into a single scalar reward by defining a trade-off a priori [118]. However, this may fail to adequately address all needs and concerns, as such a static trade-off might not be able to capture the complex preferences. To address this limitation, we explore the use of MORL for personalization within a mental health app, which allows us to explicitly account for the diverse objectives. This leads us to the following research question:

How can multi-objective reinforcement learning be used to personalize coping skill-enhancing challenges for adolescents to address needs and concerns from diverse stakeholders?

Which will be answered with the following subquestions:

- *What needs and concerns should be taken into account when developing a model to personalize coping challenges for adolescents?*
- *How can a multi-objective reinforcement learning model be designed for personalizing coping challenges for adolescents?*

- *How effective is a multi-objective reinforcement learning model for personalizing coping challenges for adolescents?*

1.2. Approach

To answer the research questions, we first identified the specific needs and concerns of adolescents, clinical experts, parents and society, and explored how these needs and concerns could be addressed in a personalization algorithm in chapter 2. Following these insights, we translated the discovered needs and concerns into design choices for our model formulation and personalization pipeline in chapter 3. We further used previously collected data from a randomized trial to aid in design decisions and to train our personalization algorithm. Additionally, this data was used to develop our simulation environment, in which users are simulated to test our personalization approach. This allowed us to evaluate how well our proposed solution could address the previously discovered needs and concerns. The results of these evaluations are presented and discussed in chapter 4. Lastly, we reflected and concluded our work by answering the research questions, discussing limitations, presenting directions for future work, and providing an ethical reflection in chapter 5.

2

Foundation

This chapter aims to answer the first sub-question, lay a foundation for multi-objective reinforcement learning, and explain the dataset used.

2.1. Needs and Concerns

In this section, we explore the diverse needs and concerns from several stakeholders and perspectives, and answer the first sub-question:

What needs and concerns should be taken into account when developing a model to personalize coping challenges?

To identify these needs and concerns, we approached the question from the perspectives of adolescents, clinical experts, and parents. Additionally, we take a societal viewpoint and discuss the ethical considerations that arise with algorithmic personalization for adolescent mHealth apps. These stakeholders were chosen because previous studies consider users (adolescents) and health care experts as stakeholders for their analysis of their perspectives on mental health apps [53, 52, 98]. Additionally, Radovic et al. [98] also considers parents and policymakers. Therefore, we considered them as well, but took the policymakers broader as society to include ethical and legal considerations.

2.1.1. Adolescents

The end users of the mental health app are adolescents aged 16-25 years old from the general population. d'Halluin et al. [31] argue that sustained engagement of adolescents in mental health applications depends on the ability of the app to meet their needs and concerns. This suggests we should have an understanding of these needs and concerns to successfully develop our personalization algorithm.

The reason behind using a mental health app is to support or improve adolescents' mental well-being. This suggests that the app should fulfill this goal and have a positive effect on users' mental health outcomes [65, 92]. Next to seeing actual improvement in well-being and mental health outcomes, it is crucial that the challenges are perceived as useful. Liverpool et al. [78] identified perceived usefulness as a major motivating factor, which suggests that users are more likely to complete a coping challenge if they think it will be useful. This aligns with the technology acceptance model [82], which argues that a user's intention to use a system depends on their belief about its usefulness and ease of use.

A further consideration is that adolescents value health applications that have engaging elements [36, 72]. Studies have found that adolescents prefer mental health applications that have fun and interactive components [49, 65], since this helps with keeping them engaged. This is in line with several technology acceptance and adoption models [112], which suggest that perceived enjoyment is an important factor that may determine engagement with technology. Such interactivity can be achieved by, for example, gamification or adding videos to the app. Additionally, young people dislike repetitiveness [127], so the app should offer a diverse variety of challenges.

A mental health app should not require too much effort or time, as this can be seen as burdensome and negatively impact the engagement of users in these apps [33]. For example, Laure [68] found that the exercises that required the most effort had the lowest completion rates. The time constraint concerns both the time required for personalization of the app (e.g. tweaking settings, or filling in questionnaires for personalization) as well as for the exercises of the app [53]. Additionally, Götzl et al. [53] and Laure et al. [69] found that young people prefer apps that encourage them to refrain from the use of mobile phones and reduce screen time. Therefore, the application should not require the users to constantly be on their phones in order for it to be effective.

Several studies investigating user needs and preferences in mental health apps indicate that users appreciate a sense of autonomy and control [5, 65, 53]. For example, Alqahtani and Orji [5] analyzed user reviews and found that users want an app where they can choose the content themselves. When looking at the specific target group of young people, Kenny, Dooley, and Fitzgerald [65] discovered that young people do not like being told what to do and that they want to be in control of how they use such apps. Moreover, Götzl et al. [53] reported that adolescents want apps to provide a variety of exercises along with an overview and orientation over possible options, so they can define their own path. These findings align with the self-determination theory [102], which identified autonomy as one of the three psychological needs for intrinsic motivation.

Another common desire is to feel connected with others and part of a group. Adolescents like to have social interaction in an app, which could be directly, by being able to send messages to peers, or indirectly, by being able to relate to others' experiences [65, 31, 78]. This social connection helps adolescents realize they are not alone in their feelings [49] and fosters a sense of belonging, which is beneficial for well-being and emotional stability [18]. Similarly to autonomy, this social relatedness is also one of the three psychological needs for intrinsic motivation, according to the self-determination theory [103].

Accessibility is also a facilitator for the success of mHealth applications. Adolescents prefer applications that can be used at any time and in any location, allowing them to engage with the app in a way that fits their routines [127]. For example, suggested exercises should take the user's current context into account and avoid requiring actions that may not be feasible, such as going outside when this is not possible. Additionally, adolescents highlight that apps should be free to use [65, 127]. Lastly, accessibility should also extend to inclusive design considerations, ensuring that the app can be used by users with physical constraints, such as color blindness or limited mobility.

When using a mental health app, young people share sensitive personal information with the application, which can raise concerns regarding privacy and safety. Kenny, Dooley, and Fitzgerald [65] reported that apps should be confidential, for example, by being password-protected, and that users should have control over their privacy, allowing them to choose what personal information is shared. Moreover, Wies, Landers, and Ienca [119] argue that privacy breaches can lead to patient mistrust and decrease the effectiveness of mHealth applications. Additionally, Liverpool et al. [78] highlighted that young people were more inclined to use apps if there was transparency. In the context of data-driven personalization, this could include transparency about how user information is used to generate recommendations. Another factor that ties closely to privacy and security is the fear of stigmatization that adolescents have. Ribanszki et al. [99] found that adolescents were cautious with using an app when there was a high risk of being seen, and Kenny, Dooley, and Fitzgerald [65] previously found the same, where users did not want others to know they were using a mental health app. This further demonstrates the need for a private, safe, and secure application.

2.1.2. Clinical experts

In addition to understanding adolescents' needs, we should also consider the perspective of healthcare professionals, as they may play a role in evaluating and validating the clinical value of health applications. While therapists are generally positive about the use of mobile health apps [53, 54], there are several considerations that should be addressed when designing and implementing such applications.

A common issue in mental health apps is low user engagement, meaning the adherence rate is low and drop out rates are high [116, 76, 12]. This can be problematic, because such applications require sufficient engagement in order to be effective [123]. Therefore, ensuring user engagement by increasing adherence and decreasing dropout rates can be essential for the success of mental health apps.

In addition to the fundamental requirement of clinical usefulness, which in our case is support-

ing adolescents in practicing coping skills, healthcare experts emphasize the necessity of knowledge transfer [53], ensuring that the skills practiced within the app effectively translate to the user's real-world environment. This means that adolescents should be able to utilize and understand the learned coping strategies in their own lives after completing the exercises.

In terms of learning coping strategies, there are several coping strategies that one can use to help deal with difficult events, such as accepting the situation or finding distraction [71, 28]. Being able to adapt to the situation and deploy these techniques flexibly can support one's ability to handle negative situations [4]. This suggests that it is beneficial to learn a diverse set of coping strategies.

Similar to adolescents, healthcare experts also have concerns about the privacy and safety of mHealth applications [54, 83]. Since mHealth apps handle personal information, there are worries about other people accessing this confidential information. Additionally, when incorporating AI into a mental health app, experts highlight the importance of ensuring transparency of data usage [53].

Moreover, experts also expressed concerns about the possible negative influences of enforcing excessive digital media use [53]. High levels of app usage should not necessary lead to effectiveness, as the ultimate goal is for users to apply learned coping strategies in their daily lives rather than remain dependent on the app. This aligns with the desire of adolescents limit their time on the phone, and complements the notion of knowledge transfer, where skills practiced within the app are carried over into real-world situations.

2.1.3. Parents

Since our study focuses on adolescents aged 16-25, the needs and concerns of parents should be taken into account as well. A scoping review by d'Halluin et al. [31] indicates parents are generally positive about mental health apps and their capabilities and features. Moreover, Butorac et al. [17] conducted a qualitative study with culturally diverse young people and parents to explore their views on engagement with a digital mental health tool and found that parents highlighted the importance of acceptance and education around mental health, but on the other hand, were worried about possible stigmatization of their child. Another key concern that they found was data being accessed by unauthorized parties. This underlines the previously identified need for privacy and safety from adolescents and clinicians.

2.1.4. Society

When developing an algorithm for a mental health app, several ethical and legal concerns arise as well. In a scoping review, Wies, Landers, and Ienca [119] found that the most frequently mentioned ethical challenges are regarding privacy, confidentiality, and security of user data. This aligns with the concerns about privacy from adolescents, clinicians, and parents and suggests we should carefully select the user data that we want to use for the algorithm, and that this data is securely stored and processed. These ethical concerns are also enforced by legal frameworks, such as the General Data Protection Regulation (GDPR) [41], which states that only strictly necessary data may be collected and that personal data must be stored and processed securely.

In addition to privacy concerns, fairness and bias mitigation should also be considered, as the algorithm should not disadvantage certain demographic groups or populations, for example, by systematically suggesting less useful challenges to girls. Such biases often originate from training a model on biased datasets [84], for instance, when certain groups are underrepresented. To mitigate biases and ensure fair treatment across all users, there are several strategies that can be employed, such as ensuring diverse and representative training data by minority oversampling [109].

Another factor that arises when integrating AI into health care is the need for interpretability and explainability [118]. This refers to the ability to interpret and explain the decisions made by the algorithm to users and stakeholders. Translating this to our goal of personalizing coping challenges, we should be able to explain why a certain challenge is suggested. This ensures transparency of the system, which in turn fosters trust among users and stakeholders [11]. The importance of this is also reflected in the EU AI Act [42], a legal framework for the development and deployment of AI systems, as it introduces requirements for AI systems regarding transparency depending on their level of risk.

Lastly, the sustainability of AI should be considered, as the development and deployment of such systems come with environmental costs [121, 24]. Training and tuning models can require large amounts of energy, which is often not from carbon-neutral sources, thereby contributing to carbon emissions. While increasing model size and complexity can improve performance, it typically also leads to a higher energy demand and, consequently, a larger environmental footprint [24]. Therefore, model

design choices should account for this trade-off between performance and environmental impact.

2.1.5. Conflicting needs and concerns

Table 2.1: Overview of needs and concerns from the perspective of adolescents, clinical experts, parents, and society.

Need / Concern	Stakeholder	Description
Usefulness	Adolescents, clinicians, parents	The application should lead to noticeable improvements and be perceived as useful.
Engaging elements	Adolescents	The app should provide fun, interactive, and varied content.
Engagement	Clinicians	Users should actively use and adhere to the application to ensure its effectiveness.
Autonomy and control	Adolescents	Users should be able to make choices and feel in control of how they use the app.
Required time and effort	Adolescents	The app should not require too much time or effort.
Social connectedness	Adolescents	The app should foster a sense of belonging and shared experience.
Privacy and safety	Adolescents, clinicians, parents, society	Personal data must be securely handled and protected from unauthorized access.
Accessibility	Adolescents	The app should be free of cost, and usable anytime and anywhere.
Knowledge transfer	Clinicians	Skills practiced in the app should generalize to real-life situations.
Transparency	Clinicians	It should be clear how data is used, and how recommendations are made.
Diversity of coping strategies	Clinicians	The application should address multiple coping techniques.
Explainability and interpretability	Society	How the algorithm makes decisions should be understandable and explainable to users and stakeholders.
Fairness and bias mitigation	Society	The app should not disadvantage specific groups.
Sustainability	Society	The development and deployment of the system should consider environmental impact and resource efficiency.

The identified needs and concerns from adolescents, clinical experts, parents, and society can be found in Table 2.1. These aforementioned needs and concerns present several possible trade-offs, where we must balance conflicts between them. For instance, adolescents do not want to spend too much time or effort on mental health exercises, but the most effective elements might be more time-consuming. Doing a diverse set of challenges is both beneficial from a clinical perspective and reduces repetitiveness, which adolescents prefer; however, a challenge that targets a different coping strategy could be more effortful for the user than a coping strategy that is already well-practiced. We also see the engagement dilemma [69]: adolescents want to reduce their screen time, but the effectiveness of the application may suffer from that. What is considered effective between the different stakeholders can also differ; a challenge that experts deem important may not be perceived as useful by users. For example, Albers, Neerincx, and Brinkman [3] found that users' usefulness beliefs did not match expert scores in a smoking cessation program. Moreover, adolescents enjoy exercises that are fun, but these might be more time-consuming. So in this case, it is difficult to determine which of these needs has more importance. Taken together, these conflicting needs and ambiguous preferences demonstrate that no single objective is sufficient to capture the complexity of the personalization problem. This makes multi-objective reinforcement learning a promising approach, as it can balance multiple, potentially conflicting objectives simultaneously.

2.2. Implications for our model

Understanding the needs and concerns from different stakeholders is only the first step in the development of our algorithm, as we must also explore how we can translate these requirements into concrete design and implementation choices. The identified needs and concerns can be addressed at different design levels of the application. Some can be directly operationalized within the personalization algorithm (e.g., as state variables or objectives), while others influence how the algorithm should be designed or constrained. Finally, some requirements concern the broader system in which the algorithm is used and cannot be addressed at the algorithmic level alone. Therefore, we categorize the needs and concerns into the following three levels: 1) model components, 2) algorithm design, and 3) system design. This categorization is visualized in Figure 2.1.

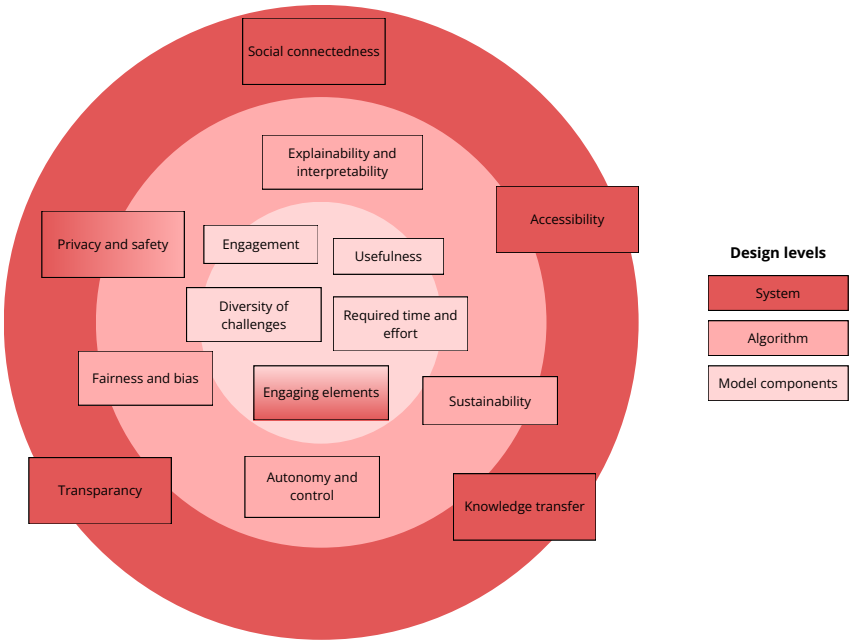


Figure 2.1: Overview of implications of needs and concerns on different design levels.

2.2.1. Model components

When developing an (MO)RL personalization algorithm, there are several components that need to be defined (see section 2.3). These include state variables, the action space, and the optimization objective(s). These components define what information the algorithm uses, which decisions it can make, and which outcomes it aims to optimize. Several of the identified needs and concerns can be directly incorporated into these components. For example, *usefulness* is typically used as an optimization objective and can be measured by an improvement in positive affect [13], perceived usefulness [6] or changes in depression and/or anxiety symptoms [108]. Interestingly, Albers, Neerincx, and Brinkman [3] used perceived usefulness as a state variable, and the usefulness according to experts as part of the reward function. This way, usefulness from multiple stakeholders' perspectives was incorporated into the algorithm. User *engagement* can also be used as an optimization objective [118]. For example, by measuring how often users use an application [58]. Ensuring the app does not require too much *time and effort* can be achieved by incorporating this in the action space, for example by selecting an activity with an appropriate difficulty [32], or by using it as an optimization objective, where we want to minimize the required time or effort. To promote the expert requirement of *diversity* of offered coping strategies, we can keep track of the completed challenges and use a diversity measure as a reward signal. This addresses adolescents' need for varied content as an *engaging element* as well.

2.2.2. Algorithm design

Algorithm design guidelines describe principles that constrain how the personalization algorithm should operate and be implemented, including the selection of inputs and outputs, and additional requirements imposed on the algorithm. For example, *autonomy and control* can be realized by letting the algorithm output a list of recommendations that the user can choose from, rather than a single suggestion [100]. Or the algorithm could allow the user to override its suggestion. To account for *safety*, we could constrain the algorithm to not give recommendations that are unsafe, for example no challenges that require going outside should be given when it is late at night. *Privacy* can be addressed on this level by carefully selecting the features that we want to use as input data, for example we might prefer tiredness data over emotion data [55], or by choosing a federated learning approach [125]. *Explainability and interpretability* can be achieved by using techniques from explainable AI (XAI) [37], such as choosing features that can be used to explain decisions or selecting an interpretable model. Mitigating *bias* and promoting *fairness* can also be addressed at this level by using diverse and representative training data and by carefully selecting input features that do not introduce or amplify existing biases. Lastly, *sustainability* can be accounted for by considering the carbon emissions of algorithms when selecting which one to use [66].

2.2.3. System design

The system design guidelines relate to the broader context in which the algorithm will be used, i.e., the whole application. For example, to foster a feeling of *social connectedness*, the application could offer the possiblity to chat with each other [35]. The need for *engaging elements* can also be fulfilled on this level. This can be done by, for example, adding gamification [35, 77], or ensuring content diversity [69]. *Privacy and safety* can be addressed on this level as well, by ensuring user data is handled securely, for example by only processing it locally on users' phones [69], or letting users choose what data is shared, simultaneously accounting for *autonomy and control*. Additionally, by giving being transparant about data collection and giving explanations of how the algorithm works, the system can ensure *transparency*. To address *accessibility*, the application should be free of cost, user-friendly, and incorporate inclusive design. Lastly, *knowledge transfer* (or 'learning transfer') can be achieved by including elements in the application that stimulate this, such as exercises that encourage the user to think about situations where skills can be applied [69].

This categorization of needs and concerns shows that not all of them can be addressed within the personalization algorithm. The system level implications regard the broader implementation of the application and fall out of the scope of algorithm development. Therefore, we will focus on the needs and concerns that can be accounted for in the model components and algorithm design level. The needs and concerns that only occur in the outer ring in Figure 2.1 will thus not be explicitly addressed in our personalization algorithm. These are transparancy, social connectedness, and knowledge transfer.

2.3. Multi-Objective Reinforcement Learning

In Multi-Objective Reinforcement Learning [59, 9, 101], an agent interacts with an environment and learns an optimal policy π^* that maps each state to an action. This can be formalized using a Multi-Objective Markov Decision Process (MOMDP), which is represented by the tuple $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mu, \mathbf{R} \rangle$, where:

- $\mathcal{S}$ is the state space,
- $\mathcal{A}$ is the action space,
- $T : S \times A \times S \rightarrow [0, 1]$ is the probabilistic transition function,
- γ is the discount factor,
- $\mu : S \rightarrow [0, 1]$ is the probability distribution over initial states,
- and $\mathbf{R} : S \times A \times S \rightarrow \mathbb{R}^m$ is a vector-valued reward function, specifying the immediate reward for each of the $m \geq 2$ objectives.

The goal of the MOMDP agent is to find a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that simultaneously maximizes the expected sum of discounted rewards J_i^π for each objective i: $\mathbf{J}^\pi = [J_1^\pi, ..., J_m^\pi]$, where J_i^π is defined as follows:

$$J_i^\pi = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, \pi(s_t), s_{t+1})\right] \tag{2.1}$$

Here, $R_i(s_t, \pi(s_t), s_{t+1})$ is the reward for objective i that is obtained from taking the action induced by the policy in state s_t, after ending up in state s_{t+1} by following the transition function. We can now also define the state-value function of a policy π as:

$$\mathbf{V}^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_{t+1} \mid s_0 = s, \pi\right] \tag{2.2}$$

where $\mathbf{r}_{t+1} = \mathbf{R}(s_t, a_t, s_{t+1})$ is the vector-valued reward received at time $t+1$. Thus, the state-value function assigns to each state $s \in \mathcal{S}$ a vector of expected discounted returns, i.e., $\mathbf{V}^\pi(s) \in \mathbb{R}^m$. In addition to the value for each state, we define the overall value of a policy π as the expectation over the initial state distribution:

$$\mathbf{V}^\pi = \mathbb{E}_{s\sim\mu}[\mathbf{V}^\pi(s)] \tag{2.3}$$

This policy-level value summarizes the expected performance of π across the entire state space and is also a vector in $\mathbb{R}^m$. In case of a single reward, the value function results in a single number, and it is easy to say that policy π is better than policy π' if $V^\pi > V^{\pi'}$. However, in the multi-objective setting we would need some utility function $u : \mathbb{R}^m \to \mathbb{R}$, that maps the multiple values to a single scalar to obtain such an ordering over policies. However, as Hayes et al. [59] thoroughly described, this function is often unknown or may change over time. Therefore, in an MOMDP, it could occur that for objective i, policy π is better (i.e. $V_i^\pi > V_i^{\pi'}$) , while for objective j, policy π' is better (i.e. $V_j^{\pi'} > V_j^\pi$). Consequently, there is no single value vector that is the best in all objectives and there are thus multiple possibly optimal value vectors $\mathbf{V}$. Therefore, we need to consider solution sets instead of a single solution as in single-objective RL and move from learning a *single* policy to learning *multiple* policies. In our problem setting, this means that for each user state, we can find a set of optimal coping challenges, each with its own optimal contributions to the objectives, instead of one challenge.

2.3.1. Comparing sets of policies

When learning sets of policies, we must be able to compare them to determine which set is better. To understand how this can be done, we explain several concepts.

Pareto-dominance and -optimality. A policy π Pareto-dominates ($\succ_P$) another policy π' when its value is at least as high in all objectives and strictly higher in at least one objective [101]:

$$\mathbf{V}^\pi \succ_P \mathbf{V}^{\pi'} \Leftrightarrow \forall i, V_i^\pi \geq V_i^{\pi'} \wedge \exists i, V_i^\pi > V_i^{\pi'} \tag{2.4}$$

If there exists no solution that dominates policy π, the policy is said to be *Pareto optimal*.

Pareto set. The Pareto set (PS) is the subset of all possible policies Π that are not Pareto dominated by any other policy in Π:

$$PS(\Pi) = \left\{\pi \in \Pi \mid \nexists \pi' \in \Pi : \mathbf{V}^{\pi'} \succ_P \mathbf{V}^\pi\right\} \tag{2.5}$$

The projection of the PS in the objective space, i.e., the values of the policies for each objective, is called the *Pareto front* (PF). This is the red line in Figure 2.2. In multi-objective optimization, the goal is often to obtain an *approximation set* of the PS, whose projection in the objective space approximates the PF.

Hypervolume. To measure the performance of an approximation set, the hypervolume metric [130] is often used. This metric provides insight into both proximity to the actual PF, as well as diversity of solutions in the approximation set. Hypervolume is measured as the volume of the region formed between each point in the approximation front and a reference point $\vec{z}_{ref}$ in the objective space (see Figure 2.2). The reference point should be chosen carefully to get meaningful insights into the quality of the approximation set. This indicator is Pareto compliant, meaning if an approximation set A dominates another set B, the hypervolume measure for set A will be strictly higher than that of B [129].

Hypervolume contribution. Following the hypervolume, the hypervolume contribution [107] is defined as the volume a certain point adds to the hypervolume. Given a solution set A, a point p, and a reference point $\vec{z}_{ref}$, the hypervolume contribution (HVC) of p to A is defined as:

$$\mathrm{HVC}(p, A, \vec{z}_{ref}) = \mathrm{HV}(A \cup \{p\}, \vec{z}_{ref}) - \mathrm{HV}(A \setminus \{p\}, \vec{z}_{ref}) \tag{2.6}$$

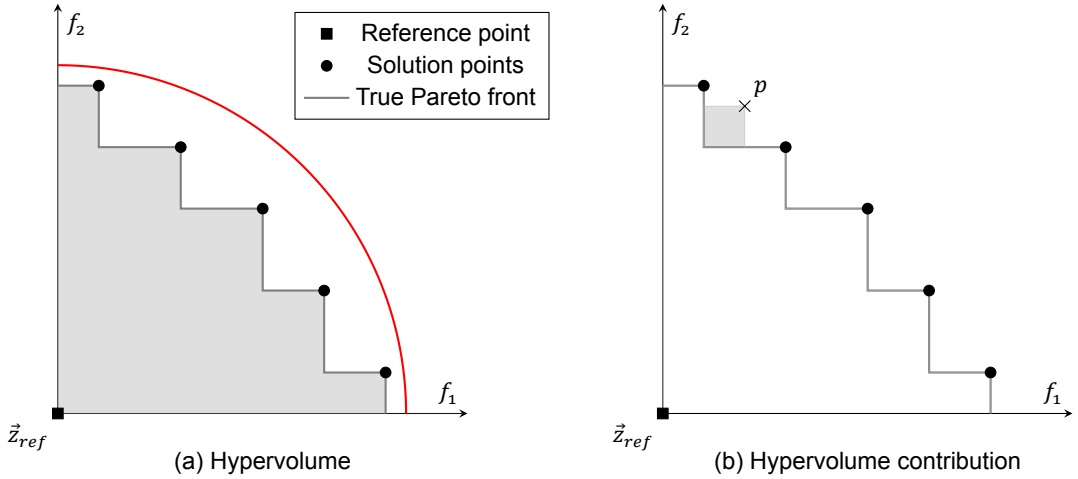


Figure 2.2: Examples of the hypervolume metric and the hypervolume contribution in a two objective maximization problem. The shaded area indicates (a) the hypervolume of the solution set, and (b) the hypervolume contribution of point x to that same set.

Expected Utility Metric. The hypervolume measure evaluates solution sets in terms of diversity and Pareto optimality, but it does not capture how well a set satisfies user preferences. To address this, Zintgraf et al. [128] introduced the expected utility metric (EUM), which evaluates the expected best utility obtainable from a solution set under uncertain user preferences.

Given a solution set A, a distribution over preference (weight) vectors Λ, and a scalarization function g, the EUM is defined as:

$$\mathrm{EUM}(A, \Lambda, g) = \mathbb{E}_{\vec{\lambda} \sim \Lambda}\left[\max_{x \in A} g(x, \vec{\lambda})\right]. \tag{2.7}$$

For each sampled preference vector $\vec{\lambda}$, the scalarization function $g(\cdot, \vec{\lambda})$ maps the multi-objective solutions to scalar utility values, and the best-performing solution for that preference vector in A is selected. The expectation then measures the average utility achievable across all possible user preferences.

2.3.2. Multi-Objective Reinforcement Learning in Healthcare

Multi-Objective Reinforcement Learning provides a framework to address problems in healthcare where there might be multiple, possibly conflicting objectives. Jalalimanesh et al. [62] used multi-objective distributed Q-learning to find the optimal radiotherapy dose while minimizing treatment time, total radiation dose, and healthy tissue damage. They took a multi-agent approach, where they used multiple agents, each optimizing one of the objectives. At the end of each iteration, the agents combine their solutions to discover Pareto optimal solutions.

Other work is done by Lizotte, Bowling, and Murphy [79], who presented an algorithm for finite-horizon fitted-Q iteration with multiple reward functions (objectives). They assumed a linear value

function approximation, meaning that the different objectives can be combined linearly to capture the true preferences. This value function takes a trade-off parameter δ, which captures the preferences for the different objectives. Using a piecewise linear representation of the value functions for the different possibilities of δ, they obtain the optimal policies under these different settings. The method is demonstrated by finding the best treatment plan for schizophrenia, which balances symptom reduction against the severity of several side effects.

2.4. Dataset

To train our model and create a simulation environment for testing, we utilized data from a randomized controlled trial study by Groot et al. [56]. This study aimed to investigate the effect of a reinforcement learning personalization algorithm on adolescents' adherence and experience with a mental health app. This algorithm was designed to determine which coping strategy category to suggest to the user. Before accessing the dataset, we preregistered our study in the Open Science Framework (OSF) [73].

2.4.1. Study Design

The study involved Dutch adolescents aged 16-25 from the general population, who interacted with a mobile app for 25 days. During this period, participants received daily challenges to enhance coping skills and Experience Sampling Method (ESM) questionnaires to capture momentary measures through the Avicenna app. The participants were assigned to one of the three between-subject conditions, which determined the level of personalization they received (random, pooled, and clustered). Additionally, there were four within-conditions which determined how the interaction between the algorithm and participants was designed: providing explanations, offering agency (possbility to select a different coping style category), a combination of both, or none of those. Before starting to use the app, participants were asked to fill in a baseline survey to assess demographics, mental health and well-being measures, and attitudes towards AI and technology. Similarly, participants, received a final survey at the end of the study to assess their mental health and well-being, and their experience with the app.

Daily Challenges. Participants received a challenge every day during the 25-day period. Before accessing their suggested challenge, they were asked a few questions about their capability, opportunity, and motivation to do a challenge (based on the COM-B behaviour change model [85]), as well as their tiredness and whether they would continue to open the challenges if they were not paid to do so. After completing a challenge, participants were asked about their experience with the challenge: how much they liked it, the difficulty, completion time, and perceived usefulness. Details on these questions can be found in Table A.1 in the Appendix.

The daily challenges were developed for the Grow It! app by Dietvorst et al. [35] and are aimed at promoting several coping strategies. They are based on cognitive behavioral therapy (CBT) and were designed in collaboration with adolescents and clinicians. The coping styles that are incorporated in the challenges are distraction, problem solving, social support, and acceptance. Furthermore, the challenges have three categories: photo challenges, quizzes, and assignments. Lastly, each challenge was also given a difficulty rating: easy, medium, or hard.

ESM questionnaires. In addition to the daily challenges, participants received five daily ESM questionnaires through the Avicenna app. These questionnaires contained questions about momentary positive and negative affect, location, and context. The morning survey included additional items about sleep quality, and the evening survey included additional questions to reflect on the events of that day and how the participant handled these.

2.4.2. Measures

The dataset consisted of the responses to the questions received before and after the challenges, the ESM questionnaire data, as well as challenge completion and participant demographics. A detailed description of the measures can be found in Table A.1 in the Appendix.

2.4.3. Data description and preprocessing

In total, there were 317 participants in the dataset from which we obtained 3471 samples that we used in our algorithm. A summary of their demographics can be found in Table 2.2. For (multi-objective) reinforcement-learning, we need samples in the form (s, a, s', r), where s is the current state, a is the action taken, s' is the next state, and r is the observed reward. To create these, we used data from the current day for s, and data from the next *active* day for s'. It could be that participants were inactive for one or more days in between these two timesteps.

While the dataset contained interactions in which users received challenges based on an RL model, we chose to still include these samples. This was done to increase the number of samples and because we assume the next state only depends on the current state (Markov property). However, data from the within-conditions where participants had the option to deviate from the suggested challenge (i.e., received agency) were excluded. The reason for this is that we believe that providing this agency may increase users' feelings of autonomy, which in turn could influence their motivation to do the challenges [102]. On the other hand, data from within-conditions where participants received explanations of the recommendations were retained. Although explanations may also affect user experience, we assumed that their influence on user behavior would be less pronounced than that of receiving agency. Therefore, these data samples were included to increase the number of samples.

Lastly, there were 14 participants who were registered under the same device identifier and exhibited unusual interaction patterns. This raised concerns among the leading researchers of the original study that these interactions may not represent independent participants. Therefore, data from these participants were excluded from our study. The numbers in Table 2.2 are exluding these 14 participants.

Data correction for photo challenges. During the data collection study, technical issues occasionally prevented participants from uploading photo challenges successfully. As a result, some photo challenges were incorrectly recorded as non-completed, and no challenge ratings were available for these entries. To quantify the extent of this issue, the final questionnaire included a question that asked participants how many photo challenges they were unable to complete due to technical difficulties. These self-reported counts were used to probabilistically correct the data, following these steps:

1. For each participant, we obtained the number of not-completed photo challenges N_{nc}, and the self-reported number of missed photo challenges due to technical issues N_{sr}.
2. To ensure that the number of corrected observations did not exceed either the observed or self-reported count, we defined the maximum number of potentially misclassified challenges as: $N_{mis} = \min(N_{nc}, N_{sr})$.
3. We estimated the correction probability for each participant: $P(\text{correction}) = \frac{N_{mis}}{N_{nc}}$
4. Per participant, each non-completed challenge was corrected with probability $P(\text{correction})$, subject to a maximum of N_{mis} corrections. For these corrected entries, the challenge was relabeled as completed and the challenge ratings (enjoyment, usefulness, difficulty, and time spent) were imputed using the mean value for the corresponding challenge among completed submissions.

Table 2.2: Participant demographics (N=317).

Characteristic	Value
Age, years	
Mean (SD)	21.7 (2.4)
Range	16-25
Gender, n (%)	
Woman	243 (76.2%)
Man	71 (22.3%)
Prefer not to say	1 (0.3%)
Other	2 (0.6%)
Education, n (%)	
Primary school	1 (0.3%)
Low	19 (6.0%)
Medium	63 (19.7%)
High	234 (73.4%)

Low = (preparatory school for) technical and vocational training; medium = (preparatory school for) professional education; high = (preparatory school for) university (preparatory track).

3

Design

This chapter describes the design of our proposed solution that answers the second sub-question:

How can a multi-objective reinforcement learning model be designed for personalizing coping challenges?

To answer this question, we designed a personalization pipeline that incorporates the needs and concerns identified in chapter 2, which come from various stakeholders, such as adolescents and experts. An overview of this pipeline is shown in Figure 3.1 and consists of three main components. First, we prepared the data from a previously collected dataset that is used for our personalization algorithm and extracted usable samples for our algorithm. This consisted of the challenges and the user interaction with these challenges, as described in section 2.4. Secondly, we designed and implemented a multi-objective reinforcement learning model that personalizes coping challenges while addressing the diverse needs and concerns identified in chapter 2. An overview of how these needs were operationalized in our design can be found in Table 3.1. This model produced a set of optimal policies that represent different trade-offs between objectives. Finally, we proposed three metapolicies, or final recommendation strategies, to determine which policy, or combination of policies from this set of optimal policies, should ultimately be presented to users.

3.1. Model components

As shown in Figure 2.1 and discussed in section 2.2, we considered the following needs and concerns on the model component level: diversity of coping strategies across completed challenges (clinical experts), engagement (clinical experts), usefulness (clinical experts and adolescents), required time/effort (adolescents), and engaging elements (adolescents). In section 2.3, we formally described MORL using a Multi-Objective Markov Decision Process (MOMDP). We modelled our problem setting using such a MOMDP, and we discuss our design choices for incorporating the needs and concerns in the action space (subsection 3.1.1), the state space (subsection 3.1.2), and the reward functions (subsection 3.1.4). Table 3.1 presents an overview of how these needs are incorporated in our design.

3.1.1. Actions

The mHealth application contains 103 coping challenges, which we grouped into a smaller number of action clusters for use in our algorithm. Estimating the effects of these challenges separately and reliably would require large amounts of data, since we want sufficient data for each (s, a)-pair. Therefore, we clustered the challenges based on average user ratings of their difficulty, fun, and perceived usefulness. The intuition behind this is that the challenges with similar user experiences are likely to result in similar transition and reward dynamics. Our MORL model thus operates over clusters of challenges rather than individual items.

We constructed these clusters using k-means clustering [1] , chosen for its simplicity and interpretability. The choice of the number of clusters was made jointly with the state space representation and is

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

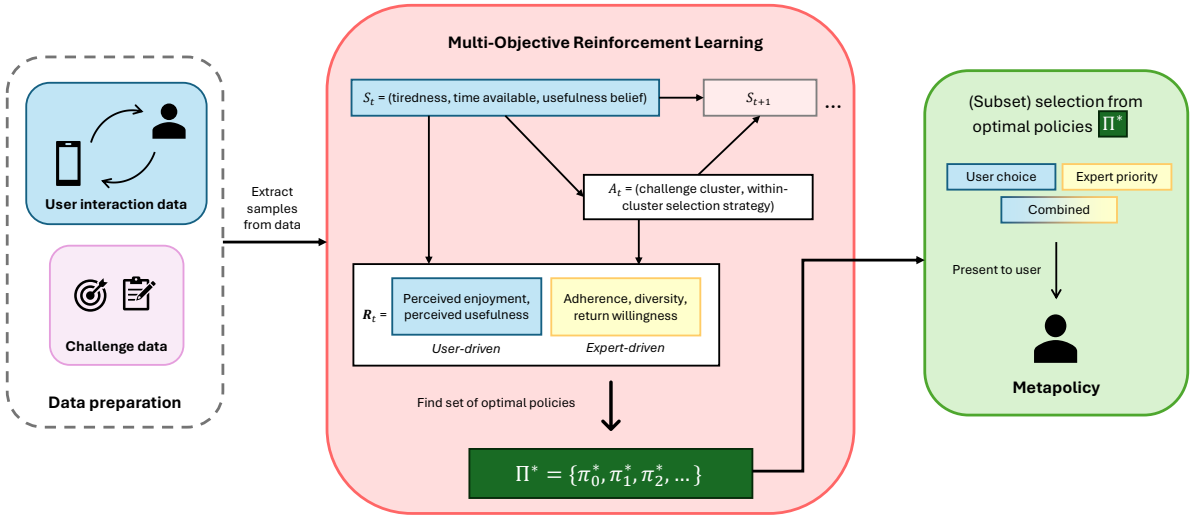
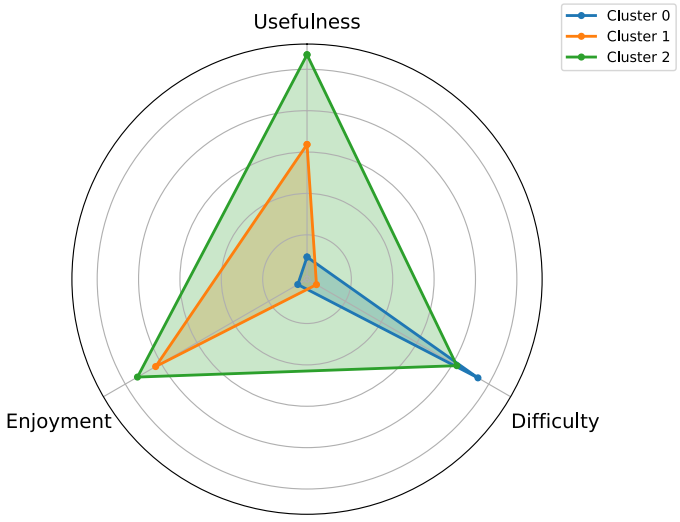


Figure 3.1: Overview of our personalization pipeline. The three main components are data collection, the multi-objective reinforcement learning model, and the final recommendation strategy.

described in subsection 3.1.2. Specifically, we selected both the state representation and the number of clusters such that the resulting (s, a)-space yields a reasonable minimum number of samples per (s, a)-pair, ensuring stable estimation of transition and reward dynamics. As mentioned in our OSF-preregistration [73], we aimed for a minimum of 5-10 samples per (s, a)-pair. The selection process explained in subsection 3.1.2 resulted in three challenge clusters. The characteristics of these clusters can be seen in Figure 3.2. The figure shows three distinguishable clusters, which we further refer to by their characterizing letter: cluster 0 (D) is difficult, cluster 1 (EU) is enjoyable and useful, and cluster 2 (EUD) is enjoyable, useful, and difficult.



	Cluster		
	0 (D)	1 (EU)	2 (EUD)
Usefulness	3.19	3.90	4.46
Difficulty	3.78	2.29	3.58
Enjoyment	4.72	6.08	6.25

Figure 3.2: Mean difficulty, enjoyment, and usefulness values for each cluster of challenges.

Within cluster selection. When using clusters of challenges, we must also select which specific challenge from that cluster to suggest to the user. To this end, we introduced a second component to

Table 3.1: Overview of how needs and concerns from different stakeholders were addressed in the design of our personalization algorithm.

Need / Concern	Translation to design
Adolescents	
Usefulness	Perceived usefulness as one of the rewards and one of the variables to cluster on.
Engaging elements	Perceived enjoyment of challenge as one of the rewards and one of the variables to cluster on. Challenges that are not completed before are prioritized to prevent suggesting the same challenge.
Autonomy and control	Providing user choice as one of the metapolicies.
Required time and effort	Time availability as one of the state features, and difficulty as one of the variables to cluster on.
Clinicians	
Usefulness	Adherence and diversity of coping strategies as reward functions.
Engagement	Adherence and return willingness as rewards.
Diversity of coping strategies	Resulting diversity of the suggested action as one of the rewards.
Society	
Explainability and interpretability	We selected interpretable state features, as well as a simple tabular model-based algorithm.
Fairness and bias mitigation	No sensitive features were used as input to the algorithm.
Sustainability	We chose a simple tabular model-based policy iteration approach over more complex deep RL methods.
All	
Privacy and safety	We selected features that feel relatively less personal and sensitive as input to our algorithm.

our action, namely the selection strategy a_s. This is a binary value, which tells the algorithm to select a challenge from the coping category that is completed the least (1) or from another category (0). This design encourages diversity in the recommended coping strategies, aligning with the expert need to equip users with a diverse coping skill set.

During the execution of the personalization algorithm, we keep track of the number of completed challenges per coping strategy category. An action a gives a cluster a_c, and within that cluster we select a challenge based on the selection mechanism a_s. To do so, we look at the counts per coping category and select the category or categories that match a_s. That is for 1, the category/categories that are practiced the least, and for 0, the other categories. From the cluster, we then randomly select a challenge from the appropriate category that has not been done before. If there are no uncompleted challenges from the appropriate categories, we randomly select a challenge from the whole cluster. Figure 3.3 gives an overview of the challenge suggestion logic embedded in our action space.

Thus, the final action is defined as a tuple:

$$a = (a_c, a_s)$$

where $a_c \in \{0, 1, 2\}$ selects a challenge cluster and $a_s \in \{0, 1\}$ determines within-cluster selection. This results in a total of six possible actions.

3.1.2. States

The state space defines the possible contexts that a user can be in and consists of several features. Since these features describe a user's personal state, they may come with ethical and regulatory implications. As described in chapter 2, one of the concerns for algorithmic personalization is privacy and the input data should thus be carefully selected. Similar to the work by Groot et al. [55], emotional data can be considered as state features for our algorithm. However, as they stated, using such data can come with several challenges, as people may not feel comfortable or have privacy concerns with sharing their emotions. Given these concerns, we chose to use alternative features that may feel less personal, as well as only features that we deem necessary to minimize the amount of required data

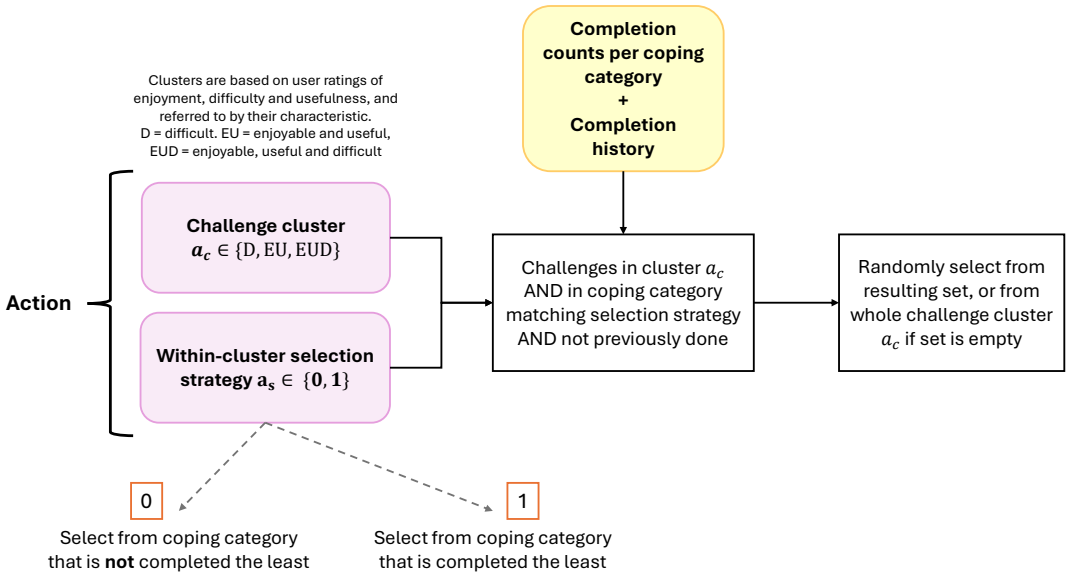


Figure 3.3: Overview of challenge suggestion given an action $a = (a_c, a_s)$.

from users. The measures that we considered were: tiredness, time available, usefulness belief, motivation, and return willingness. Details on the scales and the questions to report the measures can be found in Table A.1 in Appendix A.

COM-B elements. Since we want to predict the effects of the suggested actions in each state, the state features should be chosen so that they can accurately predict user behaviour. Albers, Neerincx, and Brinkman [3] showed that using state features derived from behaviour change theories can be effective for supporting people in smoking cessation interventions. Therefore, we adopt this notion to our mental health personalization algorithm by using the COM-B behaviour change wheel [85] as a foundation for defining our states. This theory proposes capability, opportunity, and motivation as key indicators for behaviour change. The review by Liverpool et al. [78] supports this, as they identified these three as person-specific factors that influence the engagement with DMHIs.

In the COM-B model, capability is defined as the individual's psychological and physical capacity to perform a certain action [85]. Therefore, we selected tiredness as an indicator for capability, similar to Groot et al. [55], and Albers, Neerincx, and Brinkman [3], who used energy. This was measured using a 7-point Likert scale before receiving a challenge. Opportunity is defined by factors that are outside the individual and may influence behaviour [85]. Thus, we represent it by the amount of time that users have available for doing a challenge, which is again measured on a 7-point Likert scale. Next to being one of the three elements in the COM-B model, this measure is also relevant to user needs, as we found that they do not wish to spend too much time on the mental health app [57]. Logically, this then also depends on how much time they have available in the first place, since if you do not have much time, you might prefer a shorter task. Motivation describes all the brain processes that can drive an individual to perform certain actions [85]. In our dataset, this can be represented by a user's motivation to perform a challenge, or by their usefulness belief for doing a challenge. These were both measured using a 7-point Likert scale. Another possible measure was users' return willingness (how likely they would open the app if it was not for a paid study), however, since this is not a question that would be asked during real deployment of a mental health app, we chose to not consider this measure as a state feature.

State space reduction. Using all 4 features with their 7-point scale would result in a state space of $7^4 = 2401$, which means we would require a large amount of data to reliably estimate the transition and reward dynamics in these states. Since we have a limited amount of data and we wish to have

Table 3.2: Evaluation results for different binning combinations of the feature values when using either motivation or usefulness belief as the third feature in addition to tiredness and time availability. Shown are the mean *p*-values for the difference in Q-values across different values for each feature, together with the minimum number of samples per (s, a). These are measured for different number of challenge clusters (2, 3, and 4).

		Motivation			**Usefulness belief**		
Binning combination	**Measure**	**2**	**3**	**4**	**2**	**3**	**4**
[2, 2, 2]	*p*-value	0.31	0.25	0.23	0.37	0.34	0.28
	Min. samples per (s, a)	15	11	6	14	9	5
[2, 2, 3]	*p*-value	0.28	**0.22**	0.16	0.37	0.41	0.30
	Min. samples per (s, a)	12	**7**	2	4	2	1
[2, 3, 2]	*p*-value	0.34	0.26	0.26	0.38	0.34	0.26
	Min. samples per (s, a)	5	3	2	14	9	3
[2, 3, 3]	*p*-value	0.29	0.22	0.17	0.37	0.36	0.30
	Min. samples per (s, a)	5	3	2	4	2	1
[3, 2, 2]	*p*-value	0.32	0.26	0.22	0.39	0.36	0.33
	Min. samples per (s, a)	15	11	6	14	9	2
[3, 2, 3]	*p*-value	0.29	0.23	0.19	0.39	0.39	0.34
	Min. samples per (s, a)	10	3	2	4	2	1
[3, 3, 2]	*p*-value	0.28	0.23	0.19	0.39	0.37	0.32
	Min. samples per (s, a)	5	3	1	11	7	2
[3, 3, 3]	*p*-value	0.28	0.23	0.20	0.39	0.40	0.34
	Min. samples per (s, a)	5	3	1	4	2	1

as many samples per (s, a)-pair as possible, while still retaining the relevant information in our state features, we took two measures to reduce the size of the state space: 1) selecting only one of the two motivation measures, and 2) transforming the 7-point Likert-scale features into a smaller number of bins. Additionally, we also wished to select a number of challenge clusters for our action space.

We employed a feature selection method inspired by the G-algorithm [21], similar to Albers, Neerincx, and Brinkman [3], to select one of the two motivation features, the number of values for each feature and the number of challenge clusters. For each resulting configuration, we evaluated how much the Q-values differed across feature values using ANOVA tests. Intuitively, a useful state feature should be able to differentiate between which actions are optimal, which is reflected in differing Q-values across feature values. To get an overall indication of how well the state representation is able to differentiate the values of the actions, we took the mean *p*-value across features. Lower *p*-values indicate stronger evidence that the state features are informative for dinstinguishing between actions.

Because this method requires evaluating Q-values for each configuration and thus solving the MOMDP, we employed policy iteration (see Algorithm 1) using 1000 randomly initialized weight vectors and averaged the evaluations. This allows a robust comparison of configurations across different weight vectors.

Table 3.2 shows that the motivation feature generally yields lower *p*-values than using the usefulness belief feature. Furthermore, it shows that while having 4 clusters generally results in lower *p*-values, the minimum number of samples for an (s, a)-pair is also lower. As stated in our OSF-preregistration [73], we aimed for a minimum of 5-10 samples per (s, a)-pair. Therefore, we chose to use the $[2, 2, 3]$ binning configuration with 3 challenge clusters, since this results in the lowest *p*-value while having the highest minimum number of samples per (s, a)-pair. The number of samples for each (s, a)-pair can be found in Table B.1 in Appendix B.

Formally, the state is thus:

$$\mathbf{S} = (X_{tiredness}, X_{time\_available}, X_{motivation}),$$

with

$$X_{tiredness}, X_{time\_available} \in \{0, 1\}, \qquad X_{motivation} \in \{0, 1, 2\}.$$

3.1.3. Transition function

The transition function $T(s, a, s') = P(s' \mid s, a)$ defines the probability of moving from state s to state s' after taking action a. These probabilities are estimated from the collected interaction data. For each (s, a)-pair, we count how often the next state s' occurs after taking action a in state s. Because some (s, a)-pairs may have very little data, we applied Bayesian smoothing [81] to get more stable estimates and prevent zero probabilities. For each count, we added a small amount of prior information based on the overall distribution of next states $P_{\text{global}}(s')$.

The resulting estimate of the transition function is:

$$T(s, a, s') = \frac{N(s, a, s') + \alpha P_{global}(s')}{N(s, a) + \alpha} \tag{3.1}$$

Here, α controls how strongly we rely on this prior. We set this to 1, similar to Laplace smoothing [81]. When a (s, a)-pair only has a few observations, the model leans more on the global distribution. As the number of observations increases, the influence of the prior diminishes and the estimate becomes increasingly data-driven.

3.1.4. Reward functions

In MORL, we can define multiple reward functions that represent the objectives. This is where we addressed a large part of the discovered needs and concerns. To ensure all reward signals were on the same scale, we standardized all rewards to be in $[0, 1]$. We further explain the different reward functions and how they are estimated below. In the mathematical notations, we define $n = N(s, a)$ as the number of times we took action a in state s.

Perceived usefulness. A need that was identified from all stakeholders is the usefulness of the mental well-being app. From the user perspective, we incorporated this need by using the perceived usefulness rating that is measured after completing each challenge. Let pu be the observed 7-point Likert rating, and $c \in \{0, 1\}$ be an indicator for whether the challenge was completed (1) or not (0). The normalized reward r_{pu} and its estimator are then defined as follows:

$$r_{pu}(pu, c) = \begin{cases} \frac{pu}{7}, & \text{if } c = 1 \\ 0, & \text{if } c = 0 \end{cases} \quad \hat{r}_{pu}(s, a) = \frac{1}{n} \sum_{i=1}^{n} r_{pu}(pu_i, c_i)$$

Perceived enjoyment. Another need that emerged from adolescents was their desire for fun, engaging elements. Therefore, one of the objectives was to maximize the perceived enjoyment. Let pe be the observed response to the question *"How much did you like the challenge?"* (0-10 scale), and $c \in \{0, 1\}$ be an indicator for whether the challenge was completed (1) or not (0). We defined the normalized reward r_{pe} and its estimator, similar to the perceived usefulness reward:

$$r_{pe}(pe, c) = \begin{cases} \frac{pe}{11}, & \text{if } c = 1 \\ 0, & \text{if } c = 0 \end{cases}, \quad \hat{r}_{pe}(s, a) = \frac{1}{n} \sum_{i=1}^{n} r_{pe}(fu_i)$$

Adherence. A recurring concern raised by clinical experts is low user engagement with digital health applications, reflected in low adherence and high drop-out rates [76]. Therefore, we introduce a reward signal that directly targets challenge completion. Let $c \in \{0, 1\}$ be an indicator for whether the suggested challenge was completed (1) or not (0). The reward and its estimator are then defined as follows:

$$r_{adherence}(c) = c, \quad \hat{r}_{adherence}(s, a) = \frac{1}{n} \sum_{i=1}^{n} r_{adherence}(c_i)$$

Return willingness. To further address the need to address user engagement, we included a reward that aims to increase the willingness of users to return to the application. Before receiving a challenge, adolescents are asked the question *"Currently, you are taking part in a paid study. Imagine you were not being paid for opening your daily challenge. How likely would you then have been to open your*

daily challenge?", which they can answer on a 7-point Likert scale with "Absolutely not" to "Maybe" to "Absolutely". Let rw be the score given to this question at timestep $t+1$ after receiving action a in state s. The scalarized reward for timestep t and its estimator are then defined as follows:

$$r_{rw}(rw) = \frac{rw-1}{7-1}, \quad \hat{r}_{rw}(s,a) = \frac{1}{n}\sum_{i=1}^{n} r_{rw}(rw_i)$$

Diversity. Since practicing a diverse set of coping strategies can be beneficial for learning adaptive coping, we introduce a reward for diversity. This is done by using the selection strategy in our action. Recall that 1 stands for suggesting a challenge of the least completed category, and 0 for a challenge of a different category. Since doing a challenge of the least completed category increases the diversity amongst these categories, we give a reward for that action:

$$r_{diversity}(a) = a_s$$

where $a_s \in \{0,1\}$ is the selection strategy of action a as explained in subsection 3.1.1. Note that this reward is deterministic and only depends on the action.

3.1.5. Discount factor

The discount factor $\gamma \in [0,1]$ describes how much the algorithm considers future versus immediate rewards. A discount factor of $\gamma = 0$ only optimizes for the immediate reward and does not consider future rewards at all. On the other hand, a discount factor of $\gamma = 1$ weights future rewards equally as important as immediate rewards. We set the discount factor to 0.7 to favor rewards in the near future, while still accounting for distant future values, similar to Groot et al. [55]. Because instantaneous rewards are weighted more heavily, the algorithm will prioritize challenges that meet the objectives in the near future, which may be beneficial for retention rates.

3.2. Algorithm

There are several MORL algorithms [59], and which one is suitable depends on the problem setting. In our case, we modelled our problem using a discrete state and action space, alongside stochastic transitions. Since we estimated the transition and reward functions from the gathered data, we can use model-based RL methods [86]. These have various benefits, such as data efficiency, since a learned model of the environment can be used to derive optimal policies without further interaction [86, 59]. This improved data efficiency can reduce the amount of computation required during training, and thus may contribute to a lower environmental footprint [24], which addresses the need for *sustainability*. Furthermore, the tabular and explicit transition and reward functions make it possible to inspect how policies are derived from the environment dynamics, providing more *explainability and interpretability* than black-box approaches based on deep neural networks [51].

Although model-based MORL remains relatively underexplored [59], a promising direction is MORL based on decomposition (MORL/D) [45]. This approach is inspired by evolutionary algorithms (EA) [9, 75], which operate by maintaining a set (or *population*) of candidate solutions and iteratively improving them through evolutionary operators such as selection and variation. In MORL/D, each candidate solution corresponds to a weight vector that defines a single-objective reinforcement learning problem. The multi-objective problem is thus decomposed into a collection of such subproblems, each of which can be solved independently using standard RL methods. Because these subproblems are single-objective, exact model-based methods such as policy iteration can be applied, guaranteeing convergence to the optimal policy for each weight vector. We therefore adopted the MORL/D framework [45] and modified it to a model-based setting using policy iteration.

3.2.1. Model-based MORL/D

MORL/D [45] decomposes the original multi-objective problem into a set of single-objective subproblems, each representing a different trade-off between objectives. This is achieved using a scalarization function g and an associated weight vector $\vec{\lambda}$, which maps the m-dimensional reward vector $\mathbf{R}$ to a scalar value. For simplicity, we implemented a linear scalarization function using the weighted sum:

$g(x) = \sum_{i=1}^{m} \lambda_i f_i(x)$. This also has the benefit of being interpretable, since each weight gives an indication of its importance. We further structure our method into three components (see overview in Figure 3.4): 1) initialization, 2) iterative improvement, and 3) Pareto archive maintenance. The algorithm is initialized with a set of weight vectors, which represent the single-objective RL problems that are solved independently. The population is then iteratively improved while storing non-dominated solutions in a Pareto archive. The steps are further elaborated below and a detailed description can be found in Algorithm 2.

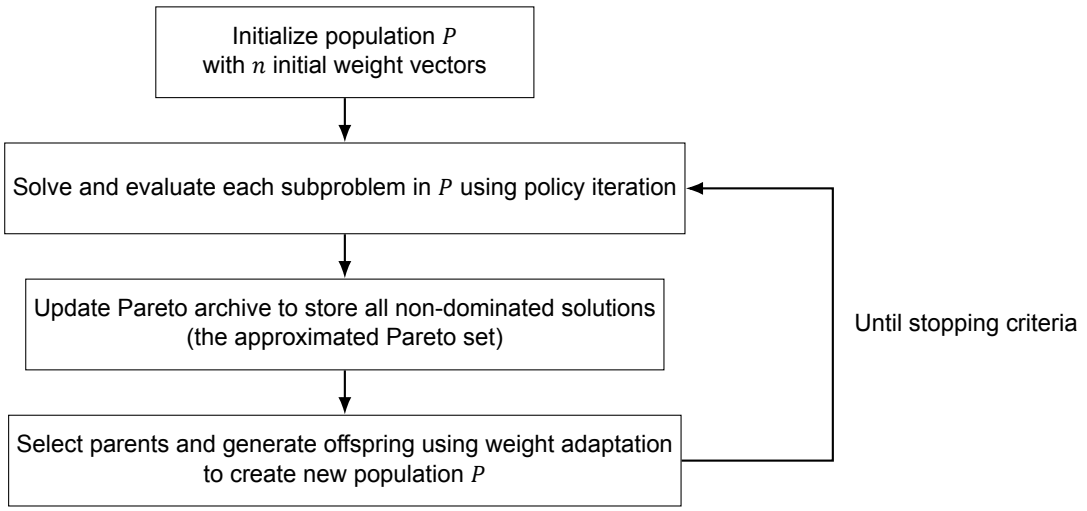


Figure 3.4: Overview of the proposed model-based MORL/D algorithm. Existing weight vectors (parents) are iteratively adapted to generate new weight vectors (offspring), after which the corresponding subproblems are solved and the Pareto archive is updated until the stopping criterion is reached.

Initialization. We initialized the population of candidate solutions with n weight vectors. Half of these are evenly distributed over the weight space [2], while the other half are sampled randomly to increase diversity. Each weight vector defines a single-objective reinforcement learning problem, which we solve using policy iteration (see Algorithm 1). This yields an initial set of policies and their value functions.

Iterative improvement. The main loop iteratively improves the population of weight vectors and corresponding policies to obtain a good approximation of the Pareto set. Each iteration consists of four steps: selection, adaptation, re-solving, and diversity maintenance.

1. *Selection.* We select $n/2$ *parent* weight vectors from the population using tournament selection. This entails randomly selecting k individuals for a tournament, and picking the one with the highest hypervolume contribution[3] until $n/2$ parents are selected. This encourages the selection of solutions that improve both quality and diversity of the Pareto approximation.

2. *Weight adaptation.* The selected parent weight vectors are then adjusted to generate *offspring* which potentially improves their performance and pushes them towards the Pareto front. This is done using Pareto Simulated Annealing (PSA) [30, 45], which compares each solution to nearby non-dominated solutions in the archive. Let $\mathbf{V}^{\pi}$ be the evaluation for policy π and $\mathbf{V}^{\pi'}$ be the evaluation for the nearest non-dominated neighbor π'. For each objective $j \in \{1, \ldots, m\}$, the weight λ_j is updated based on whether the current policy outperforms the neighbor in that specific objective:

$$\lambda_j = \begin{cases} \delta\lambda_j & \text{if } V_j^{\pi} \geq V_j^{\pi'} \\ \lambda_j/\delta & \text{if } V_j^{\pi} < V_j^{\pi'} \end{cases} \tag{3.2}$$

[2] Implemented using the Riesz s-Energy method from Blank et al. [15]

[3] We used the `moocore` Python implementation: `https://multi-objective.github.io/moocore/python/reference/generated/moocore.hv_contributions.html`

Algorithm 1 Policy Iteration

function PolicyEvaluation(π, $\mathbf{V}$, θ, *max_iters*)
  repeat
    $\Delta \leftarrow 0$
    for each state $s \in \mathcal{S}$ **do**
      $v \leftarrow \mathbf{V}(s)$
      $\mathbf{V}(s) \leftarrow \mathbf{R}(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) \mathbf{V}(s')$
      $\Delta \leftarrow \max(\Delta, |v - \mathbf{V}(s)|)$
    end for
  until $\Delta < \theta$ **or** *max_iters* reached
  return $\mathbf{V}$
end function

function PolicyIteration($\vec{\lambda}, g, \mathbf{V}_0$)
  $\pi \leftarrow$ arbitrary
  $\mathbf{V} \leftarrow \mathbf{V}_0$
  repeat
    $\mathbf{V} \leftarrow$ PolicyEvaluation(π, $\mathbf{V}, \theta$)
    for each state $s \in \mathcal{S}$ **do**
      $\pi(s) \leftarrow \arg\max_{a \in \mathcal{A}} g\Big(\vec{\lambda},\ \mathbf{R}(s,a) + \gamma \sum_{s'} P(s' \mid s, a) \mathbf{V}(s')\Big)$ # Policy improvement
    end for
  until policy converges
  return $\pi, \mathbf{V}$
end function

where $\delta > 1$ is an adaptation constant, typically around 1.05. Following the update, the weight vector $\vec{\lambda}$ is normalized to ensure $\sum_{j=1}^{m} \lambda_j = 1$. This mechanism increases the weights of objectives where the current solution already excels, effectively pushing the solution towards the Pareto front and moving away from its neighboring solutions, which enhances diversity of the approximated front [45].

3. *Re-solving and warm-starting.* Each adapted weight vector defines a new subproblem, which is solved again using policy iteration. To improve efficiency, we warm-start policy iteration using the value function of the most similar previously solved subproblem, measured by the distance between weight vectors. This is also referred to as transfer learning and can significantly reduce the number of iterations required for convergence [113, 70].

4. *Diversity maintenance.* To prevent getting stuck in local optima, we monitor the improvement in hypervolume δ_{HV} over time. If no 'significant' improvement ($\delta_{HV} < \epsilon$) is observed for several iterations, a fraction of the population is replaced with newly generated random weight vectors. These are solved using policy iteration with the same warm-start principle and integrated into the population, ensuring continued exploration of the solution space.

Pareto archive. An external archive stores all non-dominated solutions encountered during the search. After each policy evaluation step, dominated solutions are removed, and newly discovered non-dominated solutions are added. This archive provides the final approximation of the Pareto set, and thus the set of optimal policies Π^*.

3.3. Metapolicies: selecting from the Pareto optimal set

The MORL algorithm produces a Pareto archive containing a set of optimal policies Π^*, each representing a different trade-off between objectives. Recall that a policy defines an optimal challenge cluster and within-cluster selection strategy for each state. Based on the cluster and selection strategy, a specific challenge is then chosen according to the logic depicted in Figure 3.3. Our algorithm results in multiple optimal actions, and thus multiple challenges per state. This then raises the question of how

Algorithm 2 Model-based MORL/D

Input: MOMDP $\mathcal{M}$, scalarization g, population size n, iterations T, HV threshold ϵ, patience p, max resets r_{max}
Output: Pareto archive $\mathbb{PA}$

Initialization
Sample weight vectors $\Lambda = \{\vec{\lambda}_1, \dots, \vec{\lambda}_n\}$
for each $\vec{\lambda}_i \in \Lambda$ **do**
    Compute policy π_i via policy iteration; evaluate $\mathbf{V}^{\pi_i}$
    Add π_i to $\mathbb{PA}$ if non-dominated
end for
$P_0 \leftarrow \{\vec{\lambda}_i, \pi_i, \mathbf{V}^{\pi_i}\}_{i=1}^{n}$
Main Loop
for $t = 1, \dots, T$ **do**
    if HV improvement $< \epsilon$ for p consecutive iterations **then**
        if reset_count $< r_{max}$ **then**
            $R \leftarrow n/2$ random new individuals; retrain using policy iteration and update $\mathbb{PA}$
            $P_{t+1} \leftarrow$ (random half of P_t) $\cup R$
            Update reset_count
        else break
        end if
    else
        $P_{t+1} \leftarrow$ TournamentSelection(P_t, $n/2$)
        $O_t \leftarrow$ AdaptWeightsPSA(P_{t+1}); retrain O_t using policy iteration and update $\mathbb{PA}$
        $P_{t+1} \leftarrow P_{t+1} \cup O_t$
    end if
end for
return $\mathbb{PA}$

to make the final decision of which challenge or set of challenges to suggest to the user. We therefore defined three such metapolicies, which we evaluated using simulations to analyze their effects.

3.3.1. User choice

Adolescents value a sense of autonomy and control, as discussed in chapter 2. One potential way to foster this sense of autonomy and control is by allowing users to choose which exercise to complete from a set of optimal challenges rather than offering a single option. This approach is consistent with the original Grow It! [35] app, where users were also presented with three challenges, suggesting that providing multiple options aligns with the developers' initial design intentions. Providing more than one suggestion may also enhance adherence, as Fink, Newman, and Haran [46] found that increasing choice autonomy, by showing users multiple options instead of a single one, results in higher recommendation acceptance rates. This aligns with the self-determination theory by Ryan et al. [104], who argue that supporting individuals' need for autonomy can increase engagement in health behaviour change interventions and improve both mental and physical health outcomes. The meta-analysis by Carlisle et al. [19] confirms the former, as they found that providing choice can increase retention and adherence in health interventions.

While choice can enhance users' sense of autonomy and control, it must be noted that too many options can lead to *choice overload* [106]. The choice overload hypothesis states that an abundance of choice may negatively impact users' motivation and satisfaction with their choice [106]. Therefore, it is important to provide sufficient options to address users' need for autonomy, while not overloading them with too many suggestions. Since the original design of the Grow It! application included three options, and Fink, Newman, and Haran [46] saw positive effects of providing three suggestions compared to one, we chose to present users with three challenge suggestions. The number of discovered optimal policies in the Pareto archive may be very large and we thus should only select a subset of them to present to the user to prevent choice overload.

Another possible disadvantage of this strategy is that users are more likely to do the 'easy' and 'fun' challenges, which may not be the ones that most useful from a clinical perspective. However, since

adolescents prefer applications that do not take too much effort or time and enjoy elements that are engaging, we expect that providing them with these options, may keep them engaged longer. This in turn allows them to profit from the app's benefits for a longer period of time and thus results in more practise with the coping strategies.

Expected utility metric-based subset selection. To obtain a subset from the discovered Pareto optimal policies, we selected policies such that the expected utility metric (EUM) [128], given in Equation 2.7, is maximized. This is done by greedily selecting the policy from the archive that increases the EUM the most until the maximum of unique actions in any state is three. We chose the EUM because it evaluates a set of policies from the perspective of (unknown) user preferences (represented by the weight vectors) and thus results in a subset that provides high utility for a wide range of preferences.

3.3.2. Expert priority

Another method of making the final decision is selecting the policy that performs the best on the expert-driven objectives: adherence, diversity, and return willingness. This can be done by comparing the values of the policies in Π^* for the different objectives and selecting the one that yields the highest combined value for the expert-driven objectives.

3.3.3. Combination of user choice and expert priority

In addition to the two strategies described above, we could also adopt a combination of the two. For instance, we may first want to provide users with multiple options, giving them autonomy and increasing their motivation for doing the challenges, and after completing a certain number of challenges, prioritize challenges from an expert's perspective. The rationale behind this is that users may first need to gain motivation for doing the challenges before being willing to do the challenges that are more useful from an expert's point of view.

4

Evaluation

In this chapter, we will evaluate the design of our multi-objective reinforcement learning personalization algorithm and answer the last subquestion:

How effective is a multi-objective reinforcement learning model for personalizing coping challenges?

In the previous chapter, we described the design of our MORL algorithm for personalizing coping challenges within a mHealth app for adolescents. To evaluate our design, we compared several (meta)policies (section 4.1) on how they perform in user simulations across several outcomes. To this end, we set up a simulation environment (section 4.2) using the previously collected data and ran simulations to test our personalization approach. The subquestion is further split up into two analysis questions, where one assesses how well our MORL algorithm can address the discovered needs and concerns (section 4.3), and the other provides insights into the learned transition and reward functions (section 4.4).

4.1. Metapolicies and comparison policies

This section describes the different (meta)policies that we compared for suggesting coping challenges to see how they performed. As mentioned in section 3.3, we evaluated three different metapolicies, i.e., strategies to select from the set of Pareto optimal policies. We compared these against several comparison policies, as well as policies that are obtained when optimizing only one of the objectives. Table 4.1 provides an overview of the learned state-action mappings for each metapolicy and the comparison policies. The obtained single-objective policies can be found in Table D.1 in Appendix D.

4.1.1. MORL metapolicies

As mentioned in section 3.3, the output of our MORL algorithm is a set of optimal policies, and we thus aimed to evaluate multiple metapolicies, i.e., strategies of choosing from this set. We suggested and evaluated three methods: user choice, expert priority, and a combination. The evaluated (meta)policies are described below.

1. *User choice.* We selected a subset of policies such that the maximum number of different actions for any state is three, based on the expected utility metric as explained in subsection 3.3.1.
2. *Expert priority.* We selected the policy that resulted in the highest total value for the expert-driven objectives: adherence, diversity, and return willingness.
3. *Combined user choice and expert priority.* We first followed the user choice metapolicy, and after completing 12 challenges, we followed the expert priority metapolicy.

4.1.2. Comparison policies

We compared the outcomes of our metapolicies against several comparison policies: 1) random baseline, 2) equal weights, and 3) correlation-based. Each of these policies is explained below.

1. *Random baseline.* This policy simply picks a random action in each state, which represents having no personalization at all.
2. *Equal weights.* The policy that is obtained when using equal weights for each objective. This is solved using policy iteration.
3. *Correlation-based.* The policy that is found when setting the weights proportional to their correlation coefficient with challenge completion. These correlation coefficients and the corresponding weights can be found in Table C.1 in Appendix C.

The rationale behind the correlation-based policy is that we set completing challenges as a final 'goal', and each of the objectives contributes some amount to achieving that goal. We used challenge completion as the target here, with the idea that completing the challenges contributes to building adaptive coping skills.

Table 4.1: Learned policies for each (meta)policy, mapping user states to actions. States are denoted as (tiredness, time available, motivation), where tiredness and time are binary (0 = low, 1 = high) and motivation is tertiary (0 = low, 1 = medium, 2 = high). Actions are denoted as (a_c, a_s) with challenge cluster $a_c \in \{D, EU, EUD\}$ and within-cluster selection strategy $a_s \in \{0, 1\}$. MORL-UC maps each state to a set of optimal actions rather than a single action.

State	C-Corr	C-Eq	MORL-EP	MORL-UC
(0, 0, 0)	(EU,0)	(EU,1)	(EU,1)	{(EU,0), (EU,1)}
(0, 0, 1)	(EU,0)	(EU,1)	(EU,1)	{(EU,0), (EU,1), (EUD,1)}
(0, 0, 2)	(EUD,0)	(EUD,1)	(EUD,1)	{(EU,0), (EUD,0), (EUD,1)}
(0, 1, 0)	(EU,0)	(EU,1)	(EUD,1)	{(EU,0), (EU,1), (EUD,1)}
(0, 1, 1)	(EU,0)	(EU,1)	(EU,1)	{(EU,0), (EU,1), (EUD,1)}
(0, 1, 2)	(EUD,1)	(EUD,1)	(EU,1)	{(EUD,0), (EU,1), (EUD,1)}
(1, 0, 0)	(EU,0)	(EU,1)	(EU,1)	{(EU,0), (EU,1), (EUD,1)}
(1, 0, 1)	(EU,1)	(EU,1)	(EU,1)	{(EU,0), (EU,1)}
(1, 0, 2)	(EU,0)	(EU,1)	(EU,1)	{(EU,0), (EU,1), (EUD,1)}
(1, 1, 0)	(EU,1)	(EU,1)	(EU,1)	{(EU,0), (EU,1), (EUD,1)}
(1, 1, 1)	(EUD,0)	(EU,1)	(D,1)	{(EU,0), (EU,1), (EUD,1)}
(1, 1, 2)	(EU,0)	(D,1)	(D,1)	{(EU,0), (D,1), (EU,1)}

Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.
Challenge cluster abbreviations: D = Difficult, EU = Enjoyable and useful, EUD = Enjoyable, useful, and difficult.

Table 4.1 shows that the correlation-based weights (C-Corr) result in a policy that mainly suggests challenges that are characterized as enjoyable and useful. This suggests that if the goal is challenge completion, these are the types of challenges that keep users engaged. The equal weights policy and the expert priority metapolicy (MORL-EP) result in policies that only suggest challenges from a category that is the least completed. This suggests these policies account more for the diversity objective. Lastly, the user choice metapolicy (MORL-UC) recommends a diverse set of actions; however, it can be seen that the action (D,0) (difficult, not least completed category) is never suggested. This may indicate that difficult challenges are more often optimal when they are from the least completed category, perhaps giving users more incentive or motivation to complete these.

Single-objective policies. In addition to the above-mentioned (meta)policies, we also obtained the policies when only using one of the five objectives to get insights into the effects of optimizing for only one objective, for example, whether conflicts exist between objectives. Additionally, we were interested in what may be the best possible improvement achievable for each objective. These single-objective policies were derived using the same policy iteration algorithm given in 1. The obtained policies can be found in Table D.1 in Appendix D.

4.2. Simulation environment

To evaluate our personalization algorithm, we implemented a simulation environment in MO-Gymnasium [44] using the previously collected data. We used the MOMDP components that we derived for our algorithm as a basis and added some more realistic stochasticity by using distributions over outcomes instead of empirical means to mimic the real world. Figure 4.1 gives an overview of how user interaction looks in the simulation environment.

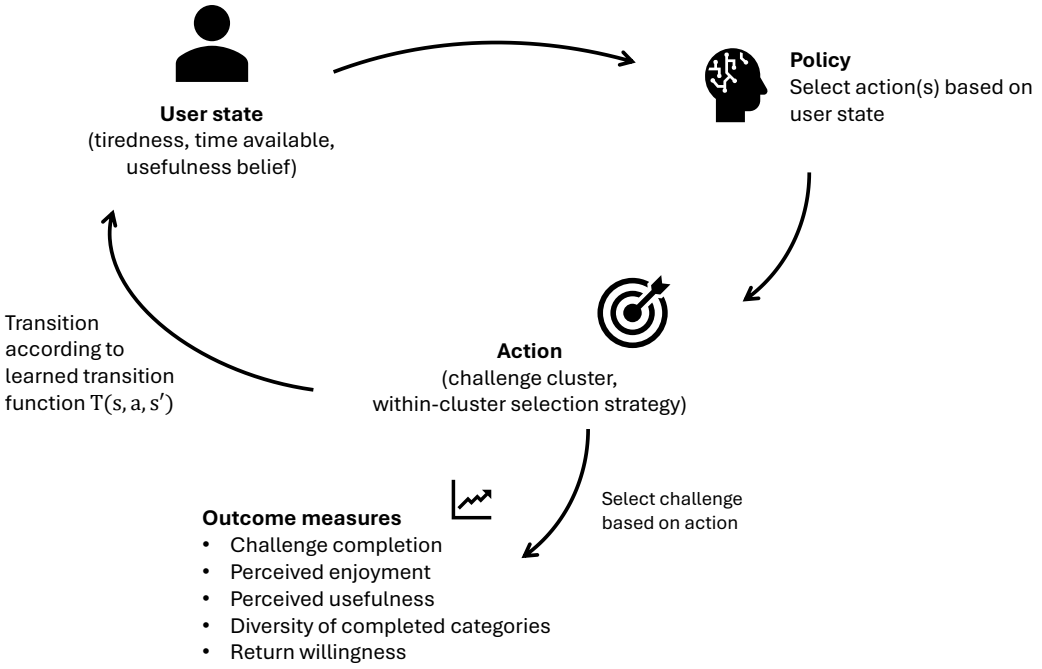


Figure 4.1: Overview of user interaction within the simulation environment. At each timestep, one or more coping challenges are recommended, and the resulting outcomes and state transitions are generated according to the learned models.

4.2.1. Outcome distributions

To examine our algorithm across various metrics, such as perceived enjoyment, perceived usefulness, and return willingness, we modeled these outcomes as stochastic variables instead of the deterministic means for each state-action pair. This is done to capture the noise and stochasticity of human behaviour. Thus, we modeled the reward functions as a probability distribution over all possible outcomes.

For an outcome o (e.g., 7-point Likert score $o \in [1..7]$), we calculated the probability of observing that outcome $P(O = o \mid s, a)$, for every state s and action a. To account for zero probabilities and data sparsity, we employed Bayesian smoothing, similar to the transition probabilities. However, instead of using the global distribution over outcomes, we used the outcome distribution per action $P(O = o \mid a)$ to fall back to. This is chosen such that when there is little data available on a certain (s, a)-pair, we assume the reward likely follows the outcome distribution induced by the action. The smoothed probability for a reward is then defined as follows:

$$P(O = o \mid s, a) = \frac{N(s, a, o) + \alpha \cdot P(o \mid a)}{N(s, a) + \alpha}$$

where $N(s, a, o)$ is the count of observed outcomes o for the pair (s, a), $N(s, a)$ is the total observations for that pair, and α is the smoothing parameter which represents the strength of our prior belief. We again set $\alpha = 1$, similar to Laplace smoothing [81]. This estimation was done for the perceived enjoyment, perceived usefulness, and return willingness measures. These were on their original scales (see Table A.1 in Appendix A), but shifted to start at 1 if needed, thus: 1-11 for perceived enjoyment, and 1-7 for perceived usefulness and return willingness. An outcome was always 0 when the challenge was not completed.

4.2.2. Completion probability

The completion probability $P(comp \mid s, a)$ defines the chance that a user in state s completes a challenge in action a. This was learned from the data similar to the outcome measures described above. In our simulations, we then simulated completion based on this probability.

4.2.3. User choice

As explained in section 3.3, one of the metapolicies that we considered is user choice, as this may increase adherence to the application. To simulate this, we increased the completion probability for this metapolicy. We performed a Bayesian paired *t*-test on the adherence of users in our dataset for conditions where they had agency (i.e., the possibility to choose another challenge) versus no agency to determine the increment in completion probability. This resulted in a mean increase of 0.031 ($\sigma = 0.013, 95\%$-HDI $= [0.006, 0.055]$) in adherence, see Table E.1 in Appendix E for further details. Therefore, we increased the completion probability for this metapolicy with a random value drawn from a normal distribution with $\mu = 0.031$ and $\sigma = 0.013$.

We further simulated which option users would choose by following a multinomial logit model, which is an often-used choice model in the literature to evaluate user choice [7, 60, 61]. In such a model, each option is assigned a utility, which represents the desirability or attractiveness of that option to the user. The probability of selecting an option is then proportional to this utility relative to the other options. We used the completion probability in the agency condition for each (s, a)-pair, denoted as $p_c(s, a)$, as the utility. This was done as the agency condition most closely reflects a setting in which users have the ability to choose from a set of challenges, and $p_c(s, a)$ thus provides an estimate of whether a challenge from action a would be chosen by the user in state s. Formally, given the completion probability $p_c(s, a)$ for each action $a \in \pi(s)$, the selection probability p_s is defined as follows:

$$p_s(s, a) = \frac{\exp[p_c(s, a)]}{\sum_{a \in \pi(s)} \exp[p_c(s, a)]}$$

The action was then chosen with respect to these probabilities, and the specific challenge was selected based on the selection mechanism (see Figure 3.3).

4.3. Addressing needs and concerns

The goal of our personalization algorithm is to address the diverse needs and concerns from different stakeholders identified in chapter 2. To assess the extent to which our multi-objective reinforcement learning (MORL) approach fulfills these needs, we answer the following analysis question:

AQ1: How well does the MORL personalization algorithm address and balance the diverse needs and concerns from different stakeholders?

For the evaluation, we focused on the needs and concerns at the 'model components' level in Figure 2.1, as these can be quantitatively assessed through our simulation environment. These were usefulness, engaging elements, required time, engagement, and diversity of coping strategies. We evaluated the (meta)policies described in section 4.1 and the evaluation proceeds in two main parts: we first show that addressing all stakeholder needs simultaneously requires navigating inherent trade-offs in subsection 4.3.1. Then we provide three more in depth exploratory analysis of adolescents' experience with completed challenges (subsection 4.3.2), the expected user retention (subsection 4.3.3), and what the diversity of coping categories looks like in practice (subsection 4.3.4).

Evaluation set-up. Using the simulation environment described in section 4.2, we ran 1000 simulations (users) for each (meta)policy over 28 timesteps (days), following the set-up by Groot et al. [55], with one or more challenges suggested at each timestep, depending on the (meta)policy. We reported on mean outcome measures across simulations, as well as the improvements over the random baseline. All policies were evaluated using the same 1000 random seeds to ensure that each policy was tested under identical simulation conditions. Further evaluation details are specified in the subsections below and an overview of the primary evaluation measures can be found in Table 4.2. Additional analyses investigating the experience of completed challenges, user retention and the practical implications of diversity are described in subsection 4.3.2, subsection 4.3.3, and subsection 4.3.4.

Table 4.2: Overview of how needs and concerns from different stakeholders in the model component level were evaluated.

Need / Concern	Evaluation measure
Adolescents	
Usefulness	Perceived usefulness outcome on a scale from 0-7. [2]
Engaging elements	Perceived enjoyment outcome on a scale from 0-11. [2]
Required time and effort	Time appropriateness measured as the ratio of mean time spent rating for challenges suggested in the high vs. low time availability state. [1]
Clinicians	
Usefulness	Seen as a combination of doing the challenges (adherence) from a broad set of coping categories (diversity).
Engagement	Adherence as the fraction of completed challenges and return willingness outcome on a scale from 0-7. [2]
Diversity of coping strategies	Shannon entropy over the coping categories of completed challenges (number in $[0, 1]$). [2]

[1] A ratio > 1 indicates the policy adapts expected challenge duration to time availability (shorter challenges when time availability is low); a value of (close to) 1 indicates little or no adaptation.
[2] Additional exploratory analyses are reported in subsection 4.3.2, subsection 4.3.3 and subsection 4.3.4.

4.3.1. Trade-offs across stakeholder needs

Since we modeled our personalization goal as a multi-objective problem, we were interested in how our approach balances these diverse objectives. We evaluated each (meta)policy across six outcome measures described in Table 4.2, reflecting the needs of both adolescents (perceived enjoyment, perceived usefulness, time appropriateness) and clinical experts (adherence, diversity, return willingness), where a higher value always indicates a better outcome. For each outcome, we reported on the relative improvement over the random baseline compared to the maximum observed improvement across all (meta)policies (including the single-objective policies) for that outcome. This is to get an idea of how well the policies perform compared to the maximum possible improvement for the objectives. These relative improvements can be seen in Figure 4.2 for the policies that only optimize a single objective and in Figure 4.3 for the other (meta)policies. The outcomes over time are shown in Figure D.3 in Appendix D.

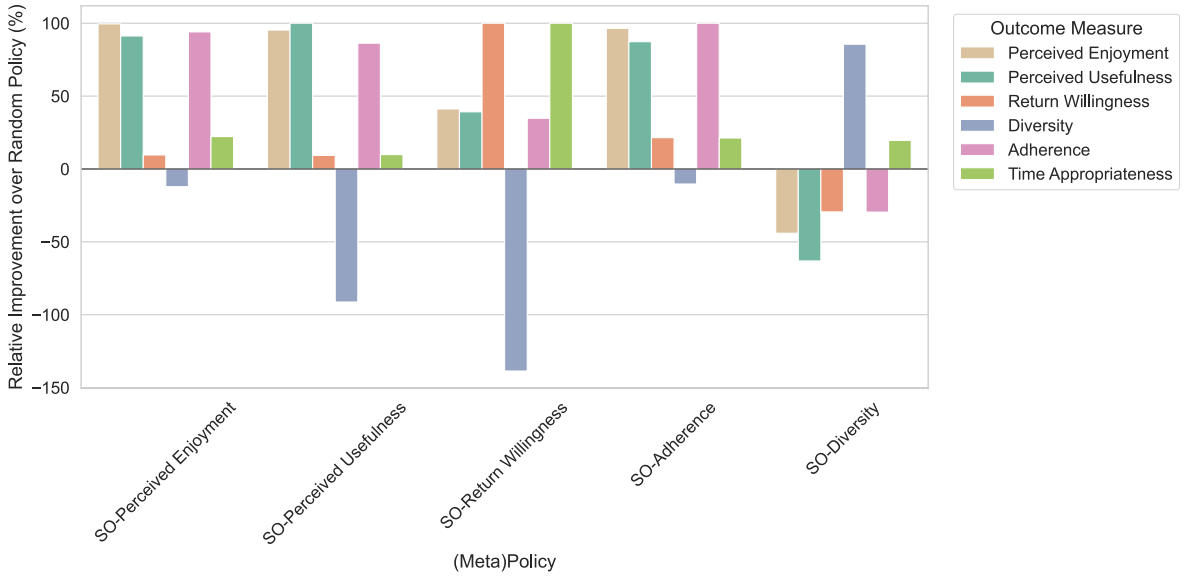


Figure 4.2: Improvements in outcomes over the random baseline policy for policies that only optimize a single objective, relative to the maximum possible improvements across all these single-objective policies and the (meta)policies.

Single-objective policies. Figure 4.2 demonstrates that no policy that only optimizes one of the rewards is able to improve over the random baseline on all outcomes. Optimizing for the adolescent-driven outcomes (perceived enjoyment and perceived usefulness) leads to decreases in diversity, whereas optimizing for diversity decreases all other outcomes. These results reveal an inherent trade-off between diversity and user engagement and satisfaction. This trade-off is not only visible *across* stakeholders, but also *within* them: optimizing for the expert-driven rewards return willingness and adherence yields a decrease in diversity as well. Overall, these findings indicate that optimizing for a single objective is insufficient for balancing all stakeholder needs. This supports the use of multi-objective reinforcement learning to address all needs and concerns and navigate the trade-offs.

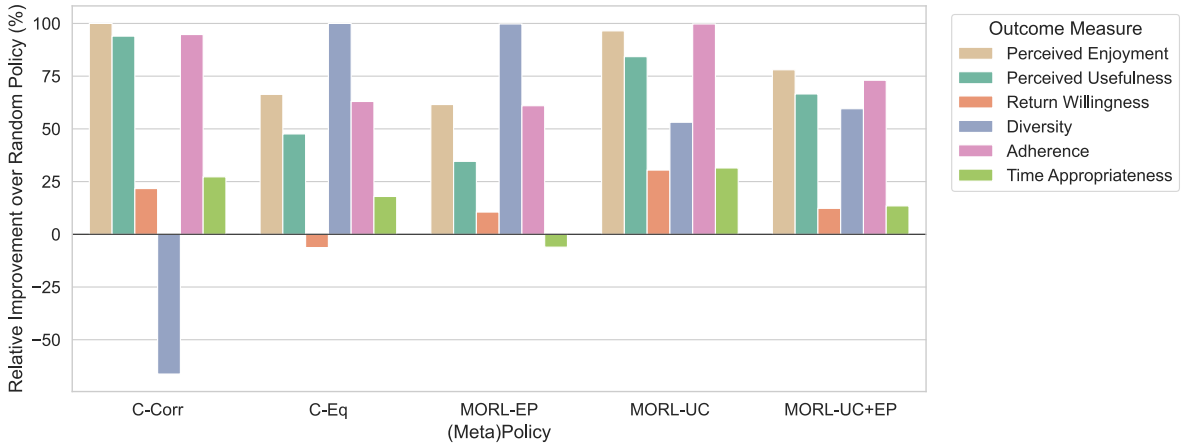


Figure 4.3: Improvements in outcomes over the random baseline policy for each (meta)policy, relative to the maximum possible improvements across all these (meta)policies and single-objective policies. Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

(Meta)Policies. Contrary to the single-objective policies, Figure 4.3 and Table 4.3 show that the user choice metapolicy (MORL-UC) and the combined metapolicy (MORL-UC+EP) improve all outcomes compared to the random baseline. It can be seen that MORL-UC outperforms MORL-UC+EP on all adolescent-driven outcomes, return willingness, and adherence. However, MORL-UC+EP achieves slightly higher diversity, as this is one of the expert-driven outcomes that it prioritizes when it switches to the expert priority metapolicy (MORL-EP). Notably, Figure D.3 in Appendix D shows this increase in diversity and a decrease in the other outcomes after switching to MORL-EP. This expert priority metapolicy prioritizes expert-driven objectives at the expense of the adolescent-driven outcomes, achieving high diversity, but failing to adapt to users' time availability and achieving lower improvements than the other (meta)policies.

The correlation-based (C-Corr) and equal weights (C-Eq) comparison policies further demonstrate the tension between diversity and other outcomes. C-Corr achieves larger improvements for all adolescent-driven outcomes, return willingness, and adherence, but decreases diversity compared to the random baseline. On the other hand, C-Eq achieves the largest increase in diversity, but performs worse on the other outcomes compared to C-Corr and even decreases return willingness.

Interestingly, C-Eq and MORL-EP achieve higher diversity than SO-Diversity, while the latter solely optimizes for diversity. This seemingly counterintuitive result may be explained by how diversity is measured: as Shannon entropy over coping categories among *completed* challenges. Therefore, higher adherence produces a more balanced distribution across categories. Since C-Eq and MORL-EP maintain better engagement and satisfaction, they achieve greater measured diversity as a byproduct. This suggests that in order to achieve meaningful diversity, user engagement must also be considered.

Taken together, these findings demonstrate that MORL-UC yields consistent improvements across all outcomes compared to the random baseline. Furthermore, the results suggest that users may be less likely to engage with challenges from a category that is completed the least. This could potentially be due to lower familiarity or comfort, or because they enjoy these challenges less. These results further emphasize that improving one objective often comes at the expense of another, highlighting

Table 4.3: Means of adolescent-driven and expert-driven outcomes for each (meta)policy across simulations, alongside the absolute improvement compared to the random baseline policy inside the parentheses. The numbers in bold are the maximum improvements.

	Adolescent-driven			Expert-driven		
(Meta)Policy	Perceived Enjoyment	Perceived Usefulness	Time Appropriateness	Return Willingness	Diversity	Adherence
Random	4.36	2.53	1.03	3.93	0.74	0.66
C-Corr	**5.67 (+29.85%)**	3.25 (+28.39%)	1.26 (+22.61%)	4.02 (+2.22%)	0.63 (-14.94%)	0.79 (+20.25%)
C-Eq	5.23 (+19.81%)	2.89 (+14.37%)	1.19 (+14.89%)	3.91 (-0.64%)	**0.90 (+22.55%)**	0.75 (+13.46%)
MORL-EP	5.16 (+18.37%)	2.79 (+10.45%)	0.98 (-5.05%)	3.98 (+1.08%)	0.90 (+22.49%)	0.75 (+13.03%)
MORL-UC	5.62 (+28.80%)	3.17 (+25.45%)	1.30 (+26.10%)	4.06 (+3.12%)	0.82 (+11.98%)	0.80 (+21.32%)
MORL-UC+EP	5.38 (+23.29%)	3.04 (+20.10%)	1.15 (+11.17%)	3.98 (+1.26%)	0.83 (+13.42%)	0.76 (+15.61%)
Single-objective policies						
SO-Adherence	5.62 (+28.80%)	3.20 (+26.37%)	1.21 (+17.62%)	4.02 (+2.20%)	0.72 (-2.32%)	**0.80 (+21.37%)**
SO-Diversity	3.79 (-13.15%)	2.05 (-19.06%)	1.20 (+16.29%)	3.81 (-3.02%)	0.88 (+19.30%)	0.62 (-6.32%)
SO-Perceived Enjoyment	5.66 (+29.72%)	3.23 (+27.56%)	1.22 (+18.47%)	3.97 (+0.99%)	0.72 (-2.73%)	0.79 (+20.10%)
SO-Perceived Usefulness	5.60 (+28.45%)	**3.29 (+30.20%)**	1.12 (+8.18%)	3.97 (+0.96%)	0.58 (-20.57%)	0.78 (+18.45%)
SO-Return Willingness	4.90 (+12.30%)	2.83 (+11.84%)	**1.89 (+83.01%)**	**4.34 (+10.25%)**	0.51 (-31.25%)	0.71 (+7.43%)

Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

the importance of considering all stakeholder needs and concerns when developing personalization algorithms.

4.3.2. Experience of completed challenges

In addition to the overall perceived enjoyment and perceived usefulness in Table 4.3, we also investigated these outcomes across the *completed* challenges, as only completed challenges provide information about how enjoyable and useful adolescents perceived the recommended challenge to be when experienced.

Table 4.4: Means of perceived enjoyment and perceived usefulness of the *completed* challenges for each (meta)policy across simulations, alongside the improvement compared to the random baseline policy inside the parentheses. The numbers in bold are the maximum improvements across these (meta)policies.

(Meta)Policy	Perceived Enjoyment (Completed)	Perceived Usefulness (Completed)
C-Corr	**7.13 (+7.99%)**	**4.09 (+6.77%)**
C-Eq	6.97 (+5.59%)	3.86 (+0.81%)
MORL-EP	6.92 (+4.73%)	3.74 (−2.28%)
MORL-UC	7.01 (+6.17%)	3.96 (+3.40%)
MORL-UC+EP	7.04 (+6.64%)	3.98 (+3.88%)
Random	6.60	3.83

Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

Results. The results in Table 4.4 show that all (meta)policies improve perceived enjoyment and perceived usefulness of completed challenges compared to the random baseline, except the expert priority metapolicy (MORL-EP). This policy decreases perceived usefulness by 2.28%, which suggests that prioritizing the expert-driven objectives comes at the cost of suggesting challenges that feel useful to adolescents. The improvement seen in Figure 4.3 may thus be attributed to the increase in adherence, and not to suggesting more useful challenges. All other (meta)policies yield improvements and are thus successful in enhancing adolescents' experience of the application by recommending more enjoyable and useful challenges.

4.3.3. Expected user retention

In subsection 4.3.1, we reported on adherence and return willingness as measures of user engagement. To further analyze engagement and explore what these numbers may mean in practice, we reported on the fraction of users remaining active over time to get an idea of user retention following the (meta)policies. For this, we assumed a user had dropped out after three consecutive days of non-completion. As this concerns the entire population of users, we performed additional simulations to get more stable estimates. We ran 20 simulations of 500 users, again with the same set of seeds for each (meta)policy. Results are shown in Figure 4.4.

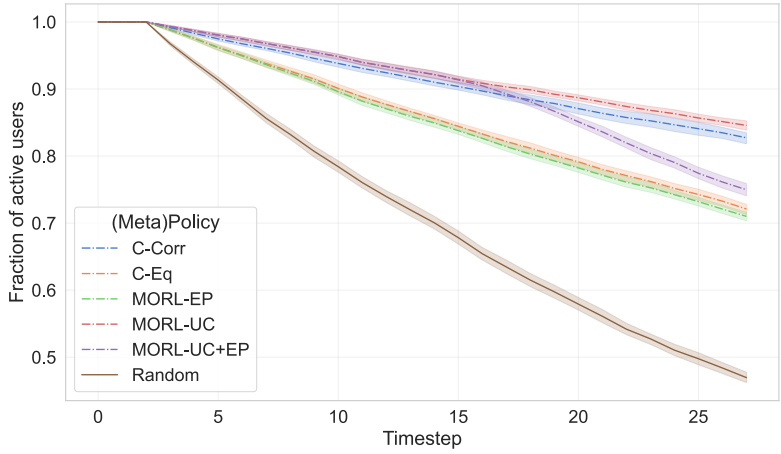


Figure 4.4: Mean fraction of active users over time for each (meta)policy, with 95% confidence intervals. Results are based on 20 simulation runs of 500 users each. Users are considered inactive after three consecutive days of non-completion. Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

Results. Figure 4.4 shows that all (meta)policies are more successful in keeping users active compared to the random baseline. When looking at the percentage of active users at the end of the simulation (i.e., after 28 timesteps), MORL-UC performs best (M=84.57%), followed closely by C-Corr (M=82.72%). C-Eq and MORL-EP achieve similar results (M=72.10% and M=71.00%, respectively), and MORL-UC+EP yields a slight drop in performance when switching from user choice to expert priority, ending with a mean of 74.98%. These results reflect the adherence numbers in subsection 4.3.1, which show that MORL-UC and C-Corr outperform C-Eq, MORL-EP, and MORL-UC+EP. Taken together, these findings indicate that personalization of coping challenges tends to increase users' willingness to continue using the system and adherence to the challenges, suggesting improved user retention.

4.3.4. Diversity of coping categories in practice

To assess whether the observed diversity in subsection 4.3.1, measured using the Shannon entropy over the coping categories of completed challenges, is meaningful in practice, we investigated how well the policies performed at reaching a target number of challenges per coping category, similar to Groot et al. [55]. We set a goal of completing at least 5 challenges per category and reported the mean fraction of achieving this goal over time. The results can be found in Figure 4.5.

Results. Interestingly, Figure 4.5 demonstrates that MORL-UC+EP achieves the highest mean fraction at the end of the simulation (0.98), followed closely by C-Eq and MORL-EP (both 0.97). This

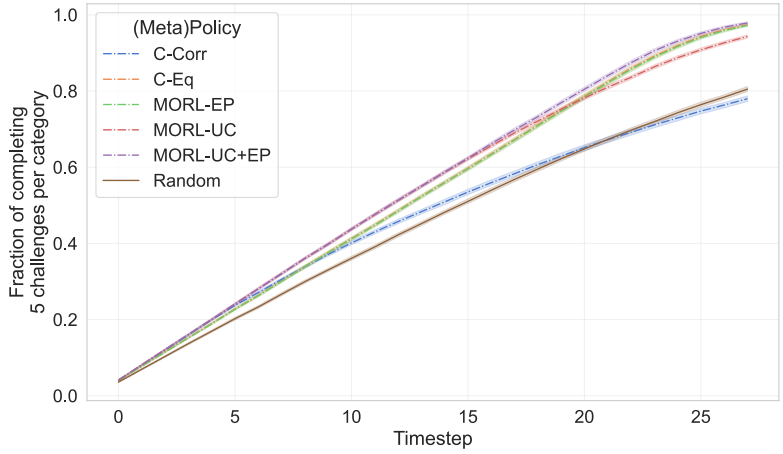


Figure 4.5: Mean fractions (with their 95% confidence interval) of completing at least 5 challenges per coping category over 28 timesteps. The lines for MORL-UC and MORL-UC+EP are completely overlapping at the beginning. Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

suggests that while C-Eq and MORL-EP achieve slightly higher entropy (see Table 4.3), this does not necessarily translate into greater success at reaching the target number of completed challenges per category, since adherence may also play a role in this. MORL-UC improves over random as well, with an average fraction of 0.94, reflecting the slightly lower Shannon entropy compared to C-Eq and MORL-EP. Lastly, C-Corr slightly decreases compared to the random baseline (0.78 compared to 0.81), which reflects the decrease in entropy shown in Table 4.3. Overall, these results indicate that all (meta)policies except C-Corr successfully engage users to complete challenges from diverse coping categories, equipping them with a broad set of coping skills.

4.4. Transition and reward functions

In our personalization approach, we estimated transition and reward functions to construct a model of user behaviour. These functions are used both by our model-based MORL algorithm for policy learning and by the simulation environment in which policies are evaluated. Consequently, the quality of the entire personalization framework depends on how accurately these functions represent the underlying user dynamics. It is therefore important to assess how accurately they capture and predict how users experience and react to the suggested actions. This evaluation provides insight into how well the learned functions capture the true environment dynamics and the extent to which they may generalize beyond the simulated setting. Thus, our second analysis question covers the evaluation of the estimated functions:

AQ2: How well can our learned transition and reward functions predict user behaviour compared to methods that do not consider state and/or action?

Evaluation method. To assess the estimated functions, we performed leave-one-user-out cross-validation (LOOCV), where the transition and reward functions are trained on all but one user and evaluated on the held-out user. We reported on the L_1-error for deterministic functions and likelihoods for stochastic functions. We further analyzed the effects of different prediction methods using Bayesian paired *t*-tests based on the Bayesian Estimation Supersedes the *t*-test by Kruschke [67][1]. For each LOOCV fold, we first computed the differences between methods per sample and then averaged these within each user to obtain a single score per user. These user-level scores were used as observations in the Bayesian analysis. We reported on the mean effect, the 95% Highest Density Interval (HDI), the posterior probability, and the probability of a ”small”, ”medium”, and ”large” effect size (set at $|0.10|$, $|0.20|$, and $|0.30|$, respectively [48]). The posterior probabilities are interpreted following the guidelines by Chechile [23] and Andraszewicz et al. [8], and for effect size, Cohen’s *d* [26] was used.

[1]We implemented the test in PyMC [1], inspired by the example in their library [110].

4.4.1. Transition dynamics

The first part of the evaluation focuses on the learned transition function, which models how users evolve over time by predicting the next state given the current state and action. To assess the predictive performance of our model, we reported on the likelihood of the observed next state under the learned transition function $P(s_{t+1} \mid s_t, a_t)$ for each state s_t. The intuition behind this is that a good transition model should assign high probability to the observed next states, reflecting that it accurately captures user dynamics. We compared our state-action transition model against two other prediction methods: 1) a uniform model that assigns equal probability to all possible next states, and 2) a model that assumes the next state is identical to the current state. The latter stay-in-state model is included because previous studies using RL for mHealth personalization have found that it is common for users to remain in their current state [2, 34], and this could thus serve as a meaningful baseline. The mean likelihoods per LOOCV fold are shown in Figure 4.6. We further performed Bayesian paired *t*-tests to compare the methods, and the posterior probabilities of our state-action model improving predictions over other methods can be found in Figure 4.7. Full summaries of the analyses can be found in Table E.2 and Table E.3 in Appendix E.

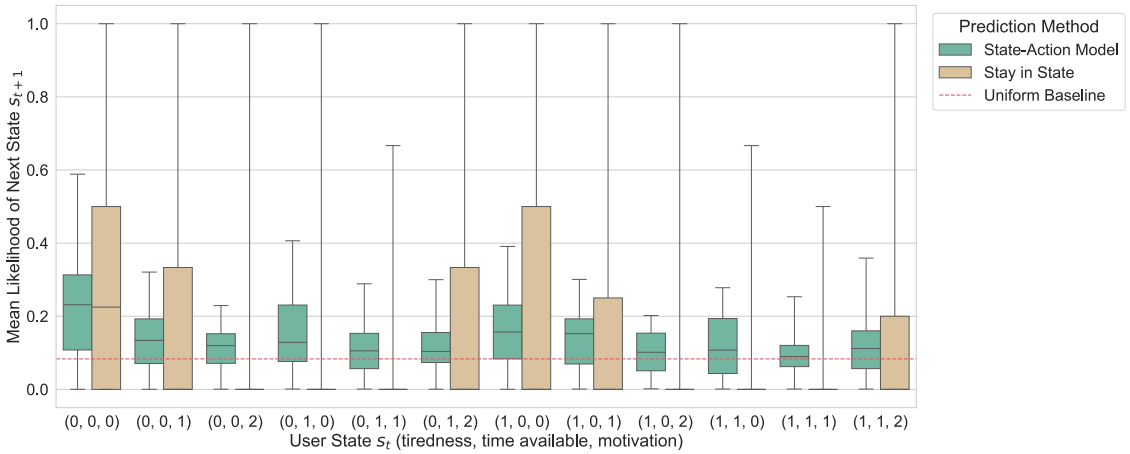


Figure 4.6: Mean likelihoods of the next state across each LOOCV fold using 1) our state-action transition model, 2) a uniform transition probability, and 3) assuming the same next state to predict the next state. The central line indicates the median, the boxes represent the interquartile range, and the whiskers show the full range. States are denoted as (tiredness, time available, motivation), where tiredness and time are binary (0 = low, 1 = high) and motivation is tertiary (0 = low, 1 = medium, 2 = high).

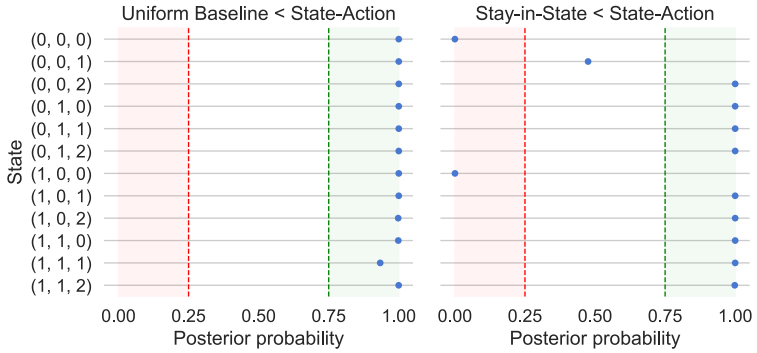


Figure 4.7: Posterior probabilities of improving the likelihoods of the next state using our state-action transition model versus 1) a uniform transition probability, 2) assuming the same next state, and 3) using an action-only model to predict the next state. The red and green regions indicate at least a casual bet against or for improvement, respectively, following the guidelines by Chechile [23] and Andraszewicz et al. [8]. States are denoted as (tiredness, time available, motivation), where tiredness and time are binary (0 = low, 1 = high) and motivation is tertiary (0 = low, 1 = medium, 2 = high).

Results. Figure 4.6 shows that the median mean likelihood assigned to the observed next state is consistently higher for the state-action model than for the uniform baseline across all states. In addition, the mean likelihood distributions of the state-action model are generally concentrated at higher

values, suggesting our transition model may improve next state predictions. The posterior probabilities in Figure 4.7 and Table E.2 confirm this, as there is at least a promising but risky bet that the state-action model improves predictions compared to the uniform baseline for all states, with posterior probabilities from 0.934 to $> .99995$. This suggests users do not transition at random, and our state-action model may be able to capture this underlying dynamic.

When comparing the state-action model with the stay-in-state model, Figure 4.6 shows that the stay-in-state model produces much more variable likelihoods, spanning the full range from 0 to 1, whereas the state-action model yields more consistent likelihoods across the LOOCV folds. This suggests that assuming users remain in the same state results in inconsistent predictive performance: while it can assign high likelihoods to some transitions, it performs poorly for others, making it a less reliable method for next state prediction. Figure 4.7 and the posterior probabilities in Table E.3 further show support for improving predictions compared to the stay-in-state baseline for almost all states (posterior probabilities varying from 0.999 to $> .99995$, all with at least a 94% large effect size). The exceptions to this are states (0,0,0), (1,0,0), and (0,0,1). In states (0,0,0) and (1,0,0), the posterior probabilities ($< .00005$ and 0.0002) even indicate near certainty against improvement. These are states where users have little time, and low motivation. This suggests that users in these states tend to stay there, and transitions to another state may be less likely. In state (0,0,1), there is no clear dominance between either prediction method (posterior probability: 0.475). In the other states, our learned transition function seems to be more likely to predict the next state, suggesting there is a more complex dynamic than staying in the current state that our state-action model aims to capture.

4.4.2. Reward prediction

The next part of our evaluation examines the performance of the learned reward functions used in our MORL algorithm. These rewards represent several needs that we identified in chapter 2, such as enjoyment, diversity, and usefulness. For the MORL algorithm, we estimate these rewards using the empirical mean for each (s, a)-pair. To evaluate the learned reward functions, we reported on the L_1-errors when 1) using our state-action model $P(r \mid s, a)$, 2) using an action-only model $P(r \mid a)$, and 3) using the global model $P(r)$ to predict the rewards.

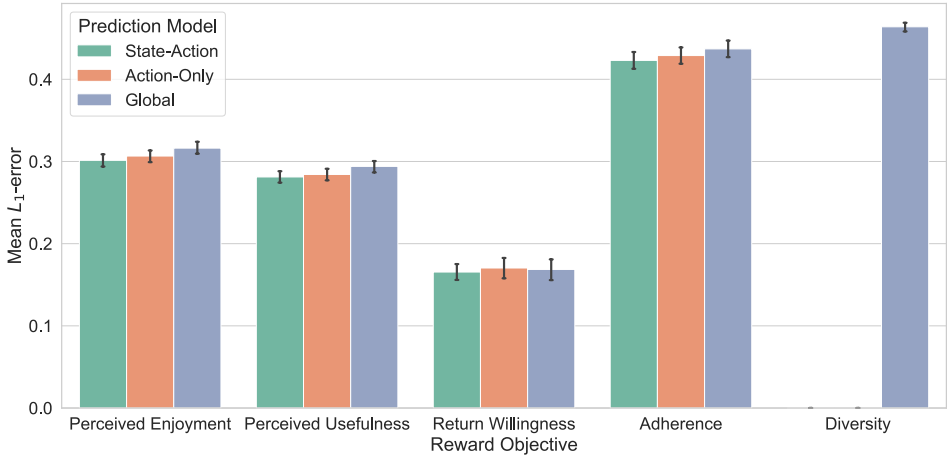


Figure 4.8: Mean L_1-errors for predicting rewards using state-action vs. action-only across LOOCV folds, alongside their 95% confidence intervals.

Results. Figure 4.8 shows that incorporating the action in reward prediction already (slightly) reduces the L_1-errors for most objectives, suggesting the action provides meaningful information for predicting these rewards. The return willingness reward is an exception to this, as the L_1-error with the global model is already quite low. This may be explained by the possibility that this is a feeling from users about the system in general, and less a result of an action as opposed to the other reward signals. For the diversity reward, both the state-action and action-only models reduce the error to 0, which is to be expected since the diversity reward is deterministic given the action.

When comparing the state-action model to the action-only model, we can see a slight decrease in errors as well in Figure 4.8. The posterior probabilities in Table 4.5 confirm this and show at least

Table 4.5: Mean reduction of L_1-errors (Δ) when using the state-action model versus the action-only model for predicting rewards, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** >)		
Objective	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
Perceived Enjoyment	0.005 $[0.003, 0.008]$	$> .99995$	Virtually certain	0.99	0.75	0.18
Perceived Usefulness	0.004 $[0.001, 0.006]$	0.996	Very strong bet	0.82	0.25	0.01
Return Willingness	0.004 $[-0.001, 0.010]$	0.951	Good bet too good to disregard	0.43	0.02	$< .01$
Adherence	0.006 $[0.003, 0.010]$	0.9998	Nearing certainty	0.96	0.56	0.08

a good bet for a reduction in L_1-errors for all objectives. For perceived enjoyment, the improvement is virtually certain (posterior probability: $> .99995$), with a 99% small- and 75% medium effect size probability. For the perceived usefulness reward, there is a very strong bet (posterior probability: 0.996) for improvement, with a 82% small effect size probability. The return willingness predictions show a good bet for improvement (posterior probability: 0.951), with a 43% small effect size probability. Lastly, the improvement for adherence is nearing certainty (posterior probability: 0.9998), with a 96% small effect size probability. Overall, these results suggest that adding state information in addition to the action improves reward prediction across all objectives.

4.4.3. Stochastic outcomes for simulations

The last part of our evaluation regards the learned outcome distributions for perceived enjoyment, perceived usefulness, return willingness, and challenge completion, which are used in the simulation environment. As explained in section 4.2, we modeled these outcome measures in our simulation environment as stochastic variables and used estimated completion probabilities to simulate challenge completions. To evaluate the predictive capabilities of our estimated outcome distributions, we performed LOOCV on the likelihoods of the outcomes. We compared the predictions of these outcomes using three different methods for obtaining the probability distributions: 1) using the state-action model $P(o \mid s, a)$, 2) using action-only $P(o \mid a)$, and 3) using the global distribution $P(o)$. For these comparisons, we performed Bayesian analysis to see whether the state-action and action-only model improve predictions over the global model, as well as whether the state-action improves predictions over the action-only model.

Table 4.6: Mean improvement Δ of outcome likelihoods when using the state-action model versus the action-only model to estimate the outcome probability distributions, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** >)		
Outcome	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
Perceived Enjoyment	-0.000 $[-0.003, 0.002]$	0.395	Not worth betting against	0.09	$< .01$	$< .01$
Perceived Usefulness	0.003 $[-0.000, 0.006]$	0.957	Good bet too good to disregard	0.52	0.06	$< .01$
Return Willingness	0.024 $[0.019, 0.028]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Completion	0.006 $[0.003, 0.010]$	$> .99995$	Virtually certain	0.97	0.57	0.08

Results. Figure 4.9 shows that both the state-action and action-only models achieve slightly higher likelihoods than using the global distributions. The posterior probabilities in Figure 4.10 confirm this

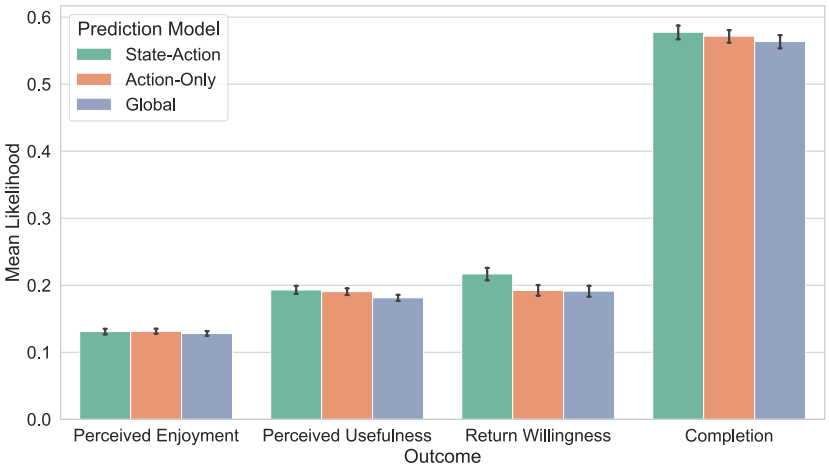


Figure 4.9: Mean likelihoods of perceived enjoyment, perceived usefulness, and return willingness across each LOOCV fold when using 1) state-action, 2) action-only, and 3) global probabilities to predict the values. The central line indicates the median, the boxes represent the interquartile range, and the whiskers show the full range.

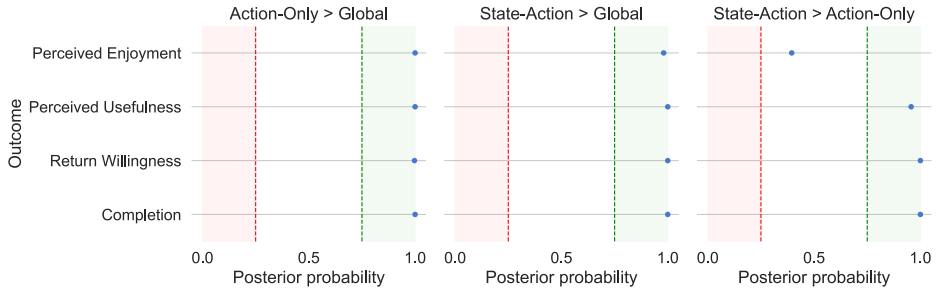


Figure 4.10: Posterior probabilities of improving the likelihoods of the outcome using our state-action probabilities versus 1) the action-only, and 2) the global probabilities to predict the values. The red and green regions indicate at least a casual bet against or for improvement, respectively, following the guidelines by Chechile [23] and Andraszewicz et al. [8].

with at least a casual bet that both action-only and state-action improve predictions for all outcomes compared to using the global distribution, with posterior probabilities ranging from 0.981 to $> .99995$. This suggests that the action already provides meaningful information for predicting these outcomes. Notably, the likelihoods also depend on the number of possible outcome values. Completion is binary, leading to higher likelihoods, whereas perceived enjoyment is on a 0-10 scale, yielding lower overall likelihoods.

It can also be seen that the state-action model improves outcome predictions compared to the action-only model. Figure 4.10 and the posterior probabilities in Table 4.6 indicate at least a good bet for improving the outcome likelihoods when adding state information for the perceived usefulness, return willingness, and completion predictions. For perceived usefulness, there is a posterior probability of 0.957, with a 59% small effect size probability, and for return willingness, the improvement is even virtually certain (posterior probability: $> .99995$) with a $> 99\%$ large effect size probability. Similarly, the posterior probability for challenge completion is $> .99995$, with a 97% small- and 57% medium effect size probability. However, for perceived enjoyment, the posterior probability is 0.395, which indicates no support against improvement ('not worth betting against'). These results indicate that the state-action model improves the prediction of perceived usefulness, return willingness and challenge completion over the action-only model, while for the perceived enjoyment this is inconclusive.

4.5. Discussion of results

This chapter presented the results of the evaluations of our multi-objective reinforcement learning personalization approach and our environment model. The results indicate that personalizing challenges using MORL can increase user engagement and enhance their experience with the challenges. Specif-

ically, our user choice policy (MORL-UC) achieves a balanced trade-off, improving all objectives compared to the random baseline (no personalization). Moreover, the estimated transition and reward functions seem to be able to improve the prediction of user behavior compared to using global estimations, suggesting our user model is able to capture these dynamics. We further discuss the two analysis questions in more detail below.

AQ1: How well does the MORL personalization algorithm address and balance the diverse needs and concerns from different stakeholders?

Table 4.7: Needs and concerns grouped by stakeholder, with changes over random baseline (++/– indicates >75% difference, +/- indicates 0-75% difference, relative to maximum possible improvement) for each evaluated (meta)policy, and an indicator of whether the (meta)policy addresses the need for autonomy (✓) or not (✗).

Need / Concern	C-Corr	C-Eq	MORL-EP	MORL-UC	MORL-UC+EP
Adolescents					
Usefulness	++	+	+	++	+
Engaging elements	++	+	+	++	++
Autonomy and control	✗	✗	✗	✓	✓/✗
Required time	+	+	-	+	+
Clinicians					
Adherence	++	+	+	++	+
Return willingness	+	-	+	+	+
Diversity of coping strategies	-	++	++	+	+

Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

The simulation results indicate that multi-objective personalization can successfully address and balance different stakeholder needs, whereas single-objective personalization fails to do so. As summarized in Table 4.7, the approach that fulfills all needs simultaneously is the user choice metapolicy (MORL-UC), which improves all outcomes compared to the random baseline and preserves user autonomy by providing choice.

The results further reveal trade-offs across stakeholder objectives: prioritizing the expert-driven outcomes comes at the cost of adapting to users' time availability. Moreover, improving diversity leads to decreased user engagement and satisfaction. The evaluation results of the single-objective policies demonstrated this tension as well and revealed that even within the set of expert-driven objectives, competition exists: optimizing for return willingness alone yields the highest return willingness outcome, but comes at the expense of diversity. This demonstrates that competing objectives exist not only *across* stakeholders, but also *within* them. Therefore, sustaining meaningful engagement requires carefully balancing objectives that may seem aligned on the surface, but conflict in practice.

The results also demonstrate the impact of the different metapolicies, i.e., how a subset is selected from the set of Pareto optimal policies that result from the MORL algorithm. As discussed in subsection 4.3.1, selecting the policy that yields the best evaluation for the expert-driven objectives (MORL-EP) does not sufficiently account for adolescent-driven outcomes, and fails to adjust to users' time availability and suggest challenges that feel useful to adolescents when completed. On the other hand, the metapolicy (MORL-UC) that uses the expected utility metric (EUM) to select a subset is more successful in achieving a balance across objectives. By selecting policies that perform well across a range of possible preference weightings, it achieves a more robust balance when these preferences are unknown. This is consistent with the theoretical motivation behind the EUM and suggests it is a promising selection strategy for MORL. These findings demonstrate that the choice of metapolicy is just as important as obtaining the Pareto optimal policies in the first place, as Pareto optimality alone does not guarantee desirable or well-balanced outcomes across stakeholder needs.

Taken together, these findings demonstrate that no single policy can be achieve the maximum improvements across all outcomes without decreasing another. Each approach achieves different trade-offs between engagement, user experience, and diversity of coping strategies. However, among the evaluated approaches, MORL-UC is the only policy that yields improvements across all outcomes

compared to the random baseline, while meeting the adolescent need for autonomy and control by providing user choice. Therefore, we believe this approach is most suitable for balancing stakeholder needs in mHealth personalization for adolescents.

AQ2: How well can our learned transition and reward functions predict user behaviour compared to methods that do not consider state and/or action?

Overall, the evaluation results suggest that incorporating state and action information leads to improved or at least comparable predictions of user behaviour compared to methods that do not consider these factors. For the transition function, the state-action model consistently outperforms the uniform baseline and provides more stable predictions than the stay-in-state model. These findings suggest that users do not transition at random, and that the state-action model is able to capture meaningful structure in user behaviour beyond what can be explained by simple assumptions about staying in the current state.

For reward and outcome prediction, adding action information already improves performance over using global estimates, and incorporating state information yields further improvements for most reward and outcome objectives. This effect is visible for all rewards, with the exception being return willingness, where the global estimate already yields relatively low prediction errors. A possible explanation is that this is a feeling about the whole system and depends less on contextual factors than other reward signals. For the stochastic outcomes, providing action information yields consistent improvements for all objectives. Extending this with the state information results in prediction improvements for perceived usefulness, return willingness and challenge completion, but for perceived enjoyment, the results were inconclusive. Although the most probable effect size for these improvements was a small effect (except return willingness, which showed 99% large effect size probability), these small improvements add up over time due to the sequential nature of our problem. Therefore, even a small effect can be consequential.

Ultimately, the results indicate that user behavior is not random but somewhat structured, and that the learned MOMDP model is able to capture this structure by jointly conditioning on both state and action. Thus, this model allows us to perform model-based RL methods and simulate user trajectories to develop and evaluate our personalization algorithm with better predictions than simpler models.

4.6. Limitations

Despite promising results for our multi-objective personalization approach being able to balance diverse needs and concerns, there are several limitations to our evaluation approach. These mainly concern the use of a simulation environment, the dataset used, and the evaluation measures.

First, the evaluation was entirely conducted in a simulated environment, which may not fully capture true human behavior and limits the extend to which the results can be generalized to real-world use. While simulation-based evaluation has several benefits, such as scalability, reproducibility and safety, and is often used in health research [16], it still relies on simplifying assumptions about user behavior and the environment. Consequently, the observed performance reflects how well the policies operate under the modeled dynamics, rather than in real-world deployment with more complex human behavior.

Another limitation is that our simulation environment was constructed from observational data that was partially collected under a different policy. The data collection study included both a random condition and a test condition in which challenges were suggested based on a reinforcement learning model. In our work, we used data from both conditions to increase the amount of data available and thus the coverage of state-action pairs, under the assumption that the Markov property holds (i.e., the next state only depends on the current state and action). However, the RL policy may have caused certain states or state-action pairs to be under- or overrepresented due to user trajectories being influenced by the policy. This may have affected the learned transition and reward functions and thus the accuracy of the evaluations.

Another data-related limitation is the fact that the data was collected in a study where users were paid to open the challenges. Although they were informed that the compensation did not depend on completion of the challenges, the monetary incentive may have increased motivation to engage with the application. Consequently, the estimated completion probabilities may be overly optimistic compared to real-world scenarios. Moreover, as explained in subsection 2.4.3, there were technical issues with

photo challenges during the data collection, which caused some data entries to be incomplete. While we aimed to correct for this as accurately as possible, the resulting estimates may still be affected by measurement errors. Consequently, some of the estimated transition and reward dynamics may be overestimated or underestimated, potentially influencing the learned policies and their evaluations.

Lastly, when evaluating the usefulness of the algorithm, we relied on proxy measures such as perceived usefulness and adherence under the assumption that completing the challenges leads to acquiring coping skills. Such proxy measures are more straightforward to obtain than long-term mental health outcomes and are therefore suitable to use in algorithms. However, these measures do not directly assess whether this observed engagement is meaningful and whether users actually improved their coping abilities.

5

Discussion and Conclusion

In this final chapter, we conclude our work by answering the research questions. We further discuss the limitations of our research and provide directions for future work. Lastly, we highlight our contributions and implications, and end with an ethical reflection on the outcomes and our research process.

5.1. Conclusions

The research question that we aimed to answer with this research was:

How can multi-objective reinforcement learning be used to personalize coping skill-enhancing challenges for adolescents to satisfy diverse needs and concerns?

Our research demonstrated that multi-objective reinforcement learning can be used to personalize coping skill-enhancing challenges for adolescents by modeling challenge recommendation as a MOMDP with multiple, potentially competing objectives that reflect the needs and concerns of different stakeholders. By incorporating adolescent context, challenge characteristics, and objectives derived from these needs, MORL can learn personalized recommendation policies that balance these outcomes. Furthermore, using a user choice metapolicy allows adolescents to select from multiple optimal recommendations, thereby addressing adolescents' desire for autonomy while maintaining personalization quality. Our simulation-based results demonstrate that this approach achieves better overall trade-offs than non-personalized or single-objective approaches and provides a practical framework for navigating tensions between stakeholder needs. Therefore, MORL offers a promising method for personalization in adolescent mHealth applications while explicitly addressing the diverse needs and concerns of adolescents, clinicians, parents, and society.

In the process of answering this main research question, we divided it into three subquestions that guided us towards answering the main question. These subquestions are discussed below to provide further insights.

RQ1: What needs and concerns should be taken into account when developing a model to personalize coping challenges?

We discovered several needs and concerns from adolescents, clinical experts, parents, and society, which have implications for different design levels of the personalization algorithm (see Figure 2.1). An overarching consideration is the *usefulness* of the application: it should be beneficial for developing coping skills, and adolescents should perceive it as useful. Adolescents further expressed the desire for an *engaging* app that offers varied and fun content. Moreover, the *required time and effort* should be appropriate to avoid burden, and adolescents should retain a sense of *autonomy and control*. From a clinical perspective, sustained *engagement* is important to ensure the effectiveness of the application. Additionally, *diversity* of coping categories supports developing a broad set of coping skills. At the societal level, several ethical and legal concerns should be considered. *Privacy and safety* are important to address, given the sensitivity of mental health data. This is in line with concerns from ado-

lescents, parents, and experts. Furthermore, *fairness and bias mitigation* are needed to ensure that the system does not disadvantage certain groups, and *explainability and interpretability* are important to ensure the decisions of the model can be understood. Lastly, the *sustainability* of the model (i.e., complexity and size) should be taken into account to limit unnecessary environmental costs. These diverse stakeholder needs and the competing nature of some of these motivate the use of multi-objective reinforcement learning, which allows us to explicitly consider trade-offs between the objectives.

RQ2: How can a multi-objective reinforcement learning model be designed for personalizing coping challenges?

Multi-objective reinforcement learning can be used to personalize coping challenges by representing challenge recommendation as a Multi-Objective Markov Decision Process (MOMDP) in which personalization is achieved by considering user context, challenge characteristics, and multiple stakeholder-driven objectives. The needs identified in RQ1 are translated into different components of the model and the algorithm choice for solving the MOMDP. Adolescents' and clinicians' desire for usefulness and engagement is operationalizaed through reward functions that capture perceived usefulness, perceived enjoyment, adherence, and return willingness. Clinical usefulness is further reflected by the requirement of developing a broad set of coping skills, which is incorporated through an objective that promotes diversity across coping categories. The need to match challenge demands with adolescents' momentary capabilities is reflected through the state representation based on COM-B elements: tiredness, time availability, and motivation. Ethical and societal concerns, including privacy, fairness, explainability, and sustainability, motivated the use of less sensitive and interpretable input features, and a simple model-based algorithm. Lastly, adolescents' desire for autonomy and control is incorporated through a metapolicy that allows users to choose from a set of optimal challenges rather than suggesting a single recommendation. Together, these design choices enable personalized challenge recommendations that align with the diverse needs of adolescents, clinicians, parents, and society.

RQ3: How effective is a multi-objective reinforcement learning model for personalizing coping challenges?

Overall, the evaluation results indicate that a multi-objective reinforcement learning model is an effective approach for personalizing coping challenges for adolescents. Unlike single-objective methods, the proposed MORL approach successfully balances competing stakeholder objectives by simultaneously improving user engagement, perceived enjoyment, perceived usefulness, adherence, diversity of coping strategies, and return willingness compared to a non-personalized baseline. In particular, the user choice metapolicy (MORL-UC) was the only (meta)policy that was able to consistently improve all outcomes compared to the non-personalized policy, while also supporting adolescents' autonomy through choice provision. The results further revealed trade-offs between adolescent-driven outcomes, such as perceived enjoyment, and expert-driven outcomes, such as diversity of coping strategies. Furthermore, we found that competing objectives exist *within* stakeholder outcomes as well: high return willingness comes at the cost of diversity of coping strategies. These tensions indicate that optimizing for a single outcome is insufficient for meeting all stakeholder needs. The findings also highlight the importance of the metapolicy used to select among the Pareto-optimal policies. Using the expected utility metric-based subset selection leads to a balance across outcomes, whereas using expert priority selection does not. In addition, the learned transition and reward functions demonstrated improved predictions of user behaviour compared to models that ignored state and/or action information, suggesting that the learned MOMDP captures meaningful user dynamics. Taken together, the results indicate that MORL is a promising approach for mHealth personalization as it enables explicit modeling and navigation of competing stakeholder objectives and can better accommodate the diverse needs and concerns than simpler personalization strategies.

5.2. Limitations

Our work comes with several limitations that should be considered when interpreting the results and conclusions. At the same time, these limitations provide several directions for future research on multi-

objective reinforcement learning and personalization in mHealth applications.

5.2.1. Uncaptured heterogeneity

A first limitation concerns the limited heterogeneity of the dataset used in this study, which consisted solely of Dutch adolescents and was largely composed of highly educated people and women (see Table 2.2). This raises questions about the generalizability of our findings to broader adolescent populations. Cross-cultural studies have found substantial differences in adolescents' coping preferences across regions, including Europe, Asia, India, and the Middle East [90, 95]. Moreover, adolescents tend to differ in their coping styles based on gender [96, 39]. As a result, the learned policies and estimated transition and reward functions may primarily reflect behavioural patterns specific to this demographic composition and may not directly transfer to more diverse populations.

Closely related to this is the fact that demographic characteristics such as gender, ethnicity, and educational background were intentionally excluded from the model for fairness considerations. While this avoids the use of sensitive features, it may also reduce the model's ability to capture meaningful heterogeneity in coping behaviour. Adolescents were found to differ in their coping tendencies based on gender [96, 39, 29], ethnicity [29], and race [22], which suggests that adolescents from different groups may benefit from different coping challenges. This is in line with Berardi et al. [14], who suggested that the design of mHealth applications should consider the gender and cultural diversity of users. Consequently, excluding these characteristics may have limited the model's ability to account for relevant individual differences in coping behaviour, potentially reducing the personalization quality across diverse adolescent populations.

5.2.2. Modeling and implementation choices

Several limitations arise from our modeling assumptions and implementation choices in our approach. First, we modeled our problem using stationary, static transition and reward functions, thereby assuming these dynamics do not change over time. However, in reality, user behavior may evolve over time due to factors such as app usage or external circumstances, which causes them to respond differently to the application [115]. Because the learned policies do not adapt to such evolving behavior, they may become suboptimal over time. This is also described by Ontanon and Zhu [91] as the "Personalization Paradox", where the personalization may influence user behavior and lose its accuracy. Additionally, the simulations may not fully capture realistic user dynamics. For example, the metapolicy that combines user choice with expert priority may perform better in practice than in our simulations, as users could become more receptive to expert-prioritized challenges after repeated exposure and practice.

Furthermore, one of our findings is that selecting from the set of Pareto optimal policies is just as important as learning these policies. However, in our work, we only tested a limited amount of such metapolicies: subset selection using the expected utility metric, and expert-priority selection. Therefore, using different methods to select the subset for users to choose from may not yield the same balance across outcomes.

5.2.3. Application in real-world settings

A further limitation that was already mentioned in section 4.6 is that the evaluation was conducted in a simulated environment rather than through tests with real users. While this provides an important first step for assessing our personalization approach in an accessible and safe manner, it remains unclear whether the identified trade-offs and results will generalize to real-world use. Future longitudinal studies with real users are therefore needed to validate these findings and determine whether balancing diverse needs and concerns translates into sustained engagement and meaningful improvements in coping skills and well-being.

Finally, this work focused specifically on the application of multi-objective reinforcement learning to the personalization of coping skill-enhancing challenges, which may limit the generalizability of the findings to other (mHealth) domains. Although some of the identified stakeholder needs and concerns, such as engagement and autonomy, are also likely to be relevant for other applications, other objectives may be domain-dependent. For example, in this work, the diversity of coping strategies was an important objective because it promotes practicing a broad set of coping skills. However, in other domains such as weight loss or physical activity promotion, such an objective may not be relevant. As a result, the observed trade-off between engagement and diversity, which is a key argument for the use of multi-objective reinforcement learning in our setting, may not be present in other mHealth

settings. In settings where such competing objectives are limited or absent, the advantages of multi-objective reinforcement learning may be less pronounced, and simpler single-objective approaches may be sufficient. Nevertheless, multi-objective reinforcement learning remains valuable for identifying and representing trade-offs and for balancing conflicting objectives when they exist.

5.3. Future research

The limitations and findings of this thesis provide several directions for future research on multi-objective reinforcement learning and personalization in mHealth applications, which are discussed below.

Conduct longitudinal user studies. A logical direction for future work is to conduct longitudinal user studies to validate whether our observed trade-offs and personalization effects also hold outside simulation environment. It would also be good to actively recruit participants from demographic groups that were less well represented in the dataset to assess whether our results can generalize to a more diverse population. Furthermore, such user studies can reveal whether the observed improvements in the proxy variables also translate to actually improving coping skills and supporting adolescents' well-being, which is the ultimate goal of the application.

Incorporate explanations. As identified in chapter 2, explainability and interpretability are some of the concerns regarding mHealth applications. While we considered this in the design process by selecting interpretable model components and by using a simple model-based algorithm instead of more complex deep learning methods, the system does not explicitly provide explanations of its recommendations to users. Future work could therefore investigate how these personalized challenge suggestions can be explained to adolescents to increase trust in the system [37, 11], an important factor for adolescents [78], and to help adolescents make more informed decisions. For example, this could be done by providing information on why the challenges are optimal compared to other challenges (based on the values for the different objectives), by learning a causal model to derive explanations [80], or by using other explainable RL methods [50].

Challenge usefulness from experts' perspective. In our model, we assumed all challenges to be equally useful in terms of learning coping skills (from an expert's perspective), while in reality, the challenges may not all equally contribute to learning these skills, as some may be more beneficial than others, and one challenge may contribute to multiple coping categories instead of only one. Therefore, additional information on the contribution of individual challenges to coping skill development could be meaningful in differentiating between the challenges and evaluating how much of each coping category a user has practiced. Therefore, future work could aim to gather such information and incorporate this in the personalization algorithm to get a better indication of the usefulness of the challenges from a clinical perspective. For example, Albers, Neerincx, and Brinkman [3] considered usefulness from an expert perspective in terms of how much each action contributed to building expert-identified competencies for preparing smoking.

Non-stationarity and partial observability. An interesting direction would be to account for the non-stationarity in adolescent behavior, where users' dynamics evolve over time, which our current approach does not capture. Future research could model the problem using a non-stationary environment and investigate whether RL methods for such environments [93] can be used to learn policies that can adjust to this dynamic user behavior. Moreover, future work could extend our MOMDP formulation to a Partially Observable Markov Decision Process (POMDP) [63], which enables the modeling of latent (unobserved) user states. While our model relies on observable indicators such as tiredness, time availability, and motivation, other factors that may influence user interaction with the app, such as stress levels, or coping abilities are difficult to observe directly. A POMDP framework could infer such latent factors from observed user behavior over time, potentially leading to more accurate user representations and more effective personalization.

Metapolicies for (subset) selection from Pareto set. Lastly, it would be interesting to test different policy selection strategies, as this has been previously identified as a challenge in multi-objective optimization [124, 126]. For example, Parisi et al. [94] propose the Local-utopia Selection Algorithm

(LUSA) to select a single policy from the Pareto set. Other possibilities would be to select on the level of actions instead of policies, which could be done by evaluating the actions in a similar manner to policies with the EUM.

5.4. Contributions and Implications

This thesis makes several contributions and has potential implications on a scientific and societal level, which are outlined below.

5.4.1. Scientific

This thesis contributes an understanding of how multi-objective reinforcement learning can be used for mHealth personalization, insights into strategies to select from Pareto optimal policies, and an overview of stakeholder needs and concerns.

Multi-objective personalization. The first scientific contribution of this thesis is the application of MORL to the personalization of mHealth applications for adolescents. This extends previous work of MORL in healthcare that addresses clinical objectives [79, 62] by explicitly accounting for user-driven outcomes. More specifically, we contribute a model-based MORL/D algorithm adapted to a discrete and stochastic setting, which is relatively underexplored compared to deep MORL methods [59]. Alongside this, we provide a simulation-based evaluation that demonstrates the practical strengths of the MORL approach over single-objective methods. Our MORL approach to generate personalized suggestions could also be applied in other settings, such as recommender systems, where multiple competing objectives must be considered as well [43, 126, 124].

Metapolicy selection. Secondly, this thesis offers insights into the consequences of different metapolicies to select from the Pareto optimal policies. Specifically, we found that subset selection based on the expected utility metric (EUM) [128] achieves a balance across outcomes, whereas expert-priority selection fails to satisfy all user needs. This demonstrates that the choice of metapolicy is as consequential as the optimization itself, and that Pareto optimality does not guarantee a good outcome in practice. These findings offer guidance for future MORL deployments on how to navigate policy selection under uncertain user preferences.

Stakeholder needs and concerns. Finally, we provide an overview of stakeholder needs and concerns that are relevant for adolescent mHealth personalization and put forward a procedure for translating these diverse needs into concrete model components. This offers additional insights into user needs, which should be understood in order to facilitate engagement in health applications [31]. Although our design is specific to this use case, the underlying design considerations may inform future studies by providing guidance on how personalization problems can be modeled in ways that align with stakeholder needs, ethical considerations, and application-specific objectives. This provides a methodology for other researchers developing MORL systems in sociotechnical contexts to better align with stakeholder needs.

5.4.2. Societal

On a societal level, this thesis has several contributions and implications for supporting adolescents' well-being, clinical decision-making, and the design and development of mHealth personalization algorithms.

Adolescents. This thesis contributes a personalization approach that better aligns with adolescents' needs and concerns, which can improve user engagement and satisfaction with the application, as our results have shown. Consequently, adolescents complete more challenges in our simulations, which in turn helps with practicing a diverse set of coping strategies. This supports adolescents by equipping them with a broad set of coping skills and may strengthen their mental resilience, aligning with the call from the World Health Organization for more support for adolescents' well-being [120].

Clinical experts. Our insights about the competing nature of stakeholder needs and concerns and the influence of the optimization objective can support experts in selecting personalization strategies that align with clinical values. More broadly, the findings in this thesis could inform clinical decision-making beyond mHealth: the modeling of competing objectives mirrors the kinds of trade-offs clinicians may face when tailoring support strategies to individual clients, and making these trade-offs transparent and quantifiable may support more informed decision-making in practice.

Designers and developers. Lastly, the findings in this thesis can aid the design and development of mHealth personalization algorithms to better address diverse stakeholder needs. Our findings suggest that explicitly modeling and optimizing for these needs using MORL can achieve a balance across the different objectives. Therefore, we provide a two-fold recommendation for future developers: 1) identify and understand relevant needs and concerns from different stakeholders, and 2) adopt a multi-objective approach to model personalization rather than collapsing competing needs into a single reward signal, as this may overlook important stakeholder values. Finally, this thesis provides a practical implementation of a personalization algorithm for the Grow It! application that can be incorporated by the developers of the app.

5.5. Ethical reflection

The increasing use of AI brings responsibility to researchers and developers to ensure alignment with human values and ethical standards to mitigate possible risks. Therefore, we discuss some ethical implications that should be considered following our research, and end with a reflection on our research process. The discussed concerns are not exhaustive, but aim to provide an idea of the possible risks.

5.5.1. Ethical implications

Privacy and security. Personalization relies on collecting and processing user data, which introduces privacy and security risks. Although we deliberately used relatively non-sensitive features, such as self-reported tiredness, time availability, and motivation, adolescents and their parents or caregivers may still have concerns regarding how personal information is collected, stored, and translated into recommendations. Moreover, data related to mental health can be perceived as highly personal regardless of its relative sensitivity. Therefore, during further developments and deployments, the design choices and their impacts should be continuously monitored, as suggested by Groot et al. [55], to ensure alignment with privacy concerns. Furthermore, users could be given the option to choose whether they want a personalized experience, and thus whether their data is used, as suggested by Götzl et al. [53] as 'personalized personalization'. Lastly, to increase transparency and facilitate trust in the application, users should be well-informed about how their data is used, for example, through a privacy policy [5].

Inequal access. While increased accessibility is one of the benefits of mental health applications, there are potential issues related to inequal access to such applications [119]. These mHealth applications may not be accessible to all adolescents, as it depends on factors such as smartphone ownership, digital literacy, and internet access, which may vary across socioeconomic contexts [105]. As a result, certain groups of adolescents may not be able to receive personalized well-being support, even if they are among those who could benefit most. Our work could unintentionally reinforce such existing disparities, as personalization enhances the quality of support for users who are already able to engage with the technology, while providing no benefit to those without access.

5.5.2. Responsible research

Reproducibility and replicability. In the interest of open science, defined as "the process of making the content and process of producing evidence and claims transparent and accessible to others" by Munafò et al. [87], we followed several principles to ensure transparency and reliability of our research. Firstly, we pre-registered our study in the Open Science Framework (OSF) prior to accessing the dataset [73]. This was done to provide a transparent record of the planned analyses and to prevent reporting bias. Furthermore, all implementation and analysis code is made publicly available on the 4TU Data Research Repository [74], and the dataset used in this study will be released upon publication. Taken together, these measures enable reproducibility and support replicability, allowing researchers to verify and build upon our findings.

Use of generative AI. During this thesis, generative AI tools (ChatGPT and Gemini) were used for three main purposes: generating illustration ideas, assisting with coding, and supporting the writing process. For all purposes, the content of the AI's suggestions were critically evaluated and validated before incorporating them.

For coding-related tasks, AI tools were primarily used to help with plotting figures and debugging code. No research data was shared with these tools during this process. Generative AI was also used to help with generating ideas to illustrate concepts in this thesis. Specifically, we prompted AI to brainstorm ideas for the illustrations of the design pipeline (Figure 3.1) and the overview of stakeholder needs and concerns (Figure 2.1). The generated suggestions only served as inspiration, and the final figures were created on our own. In addition to this, ChatGPT was used to generate the initial cover illustration of this thesis, which we further refined ourselves. Lastly, AI tools were used to support the writing and formatting process of this thesis. This included assistance with improving sentence clarity, refining the structure of paragraphs, creating tables, and formatting. Furthermore, we consulted AI to help with creating initial drafts for the LaTeX figures in Figure 3.4 and Figure 2.2, after which we refined and edited them ourselves.

Bibliography

[1] O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, M. Osthege, R. Vieira, T. Wiecki, and R. Zinkov. “PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python”. In: *PeerJ Computer Science*, *9*, 2023. doi: 10.7717/peerj-cs.1516.

[2] N. Albers, M. A. Neerincx, and W.-P. Brinkman. *Persuading to Prepare for Quitting Smoking with a Virtual Coach: Using States and User Characteristics to Predict Behavior*. 2023.

[3] N. Albers, M. A. Neerincx, and W.-P. Brinkman. “Reinforcement learning for proposing smoking cessation activities that build competencies: Combining two worldviews in a virtual coach”. In: *BMC Medical Informatics and Decision Making*, *25*, 2025, p. 370. doi: 10.1186/s12911-025-03164-8.

[4] A. Aldao, G. Sheppes, and J. J. Gross. “Emotion Regulation Flexibility”. In: *Cognitive Therapy and Research*, *39*, 2015, pp. 263–278. doi: 10.1007/s10608-014-9662-4.

[5] F. Alqahtani and R. Orji. “Insights from user reviews to improve mental health apps”. In: *Health Informatics Journal*, *26*, 2020, pp. 2042–2066. doi: 10.1177/1460458219896492.

[6] M. K. Ameko, M. L. Beltzer, L. Cai, M. Boukhechba, B. A. Teachman, and L. E. Barnes. “Offline Contextual Multi-armed Bandits for Mobile Health Interventions: A Case Study on Emotion Regulation”. In: *Fourteenth ACM Conference on Recommender Systems*. Virtual Event Brazil: ACM, 2020, pp. 249–258. doi: 10.1145/3383313.3412244.

[7] A. Anas. “Discrete choice theory, information theory and the multinomial logit and gravity models”. In: *Transportation Research Part B: Methodological*, *17*, 1983, pp. 13–23. doi: 10.1016/0191-2615(83)90023-1.

[8] S. Andraszewicz, B. Scheibehenne, J. Rieskamp, R. Grasman, J. Verhagen, and E.-J. Wagenmakers. “An Introduction to Bayesian Hypothesis Testing for Management Research”. In: *Journal of Management*, *41*, 2015, pp. 521–543. doi: 10.1177/0149206314560412.

[9] H. Bai, R. Cheng, and Y. Jin. “Evolutionary Reinforcement Learning: A Survey”. In: *Intelligent Computing*, *2*, 2023, p. 0025. doi: 10.34133/icomputing.0025.

[10] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard. “Mental Health Smartphone Apps: Review and Evidence-Based Recommendations for Future Developments”. In: *JMIR Mental Health*, *3*, 2016, e4984. doi: 10.2196/mental.4984.

[11] E. Barnes and J. Hutson. “Navigating the Complexities of AI: The Critical Role of Interpretability and Explainability in Ensuring Transparency and Trust”. In: *International Journal of Multidisciplinary and Current Educational Research*, *6*, 2024.

[12] A. Baumel, F. Muench, S. Edan, and J. M. Kane. “Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis”. In: *Journal of Medical Internet Research*, *21*, 2019, e14567. doi: 10.2196/14567.

[13] M. L. Beltzer, M. K. Ameko, K. E. Daniel, A. R. Daros, M. Boukhechba, L. E. Barnes, and B. A. Teachman. “Building an emotion regulation recommender algorithm for socially anxious individuals using contextual bandits”. In: *British Journal of Clinical Psychology*, *61*, 2022, pp. 51–72. doi: 10.1111/bjc.12282.

[14] C. Berardi, M. Antonini, Z. Jordan, H. Wechtler, F. Paolucci, and M. Hinwood. “Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework”. In: *BMC Health Services Research*, *24*, 2024, p. 243. doi: 10.1186/s12913-023-10536-1.

[15] J. Blank, K. Deb, Y. Dhebar, S. Bandaru, and H. Seada. “Generating Well-Spaced Points on a Unit Simplex for Evolutionary Many-Objective Optimization”. In: *IEEE Transactions on Evolutionary Computation*, In press.

[16] A.-L. Boulesteix, R. H. Groenwold, M. Abrahamowicz, H. Binder, M. Briel, R. Hornung, T. P. Morris, J. Rahnenführer, and W. Sauerbrei. “Introduction to statistical simulations in health research”. In: *BMJ Open*, *10*, 2020, e039921. doi: 10.1136/bmjopen-2020-039921.

[17] I. Butorac, R. McNaney, J. P. Seguin, P. Olivier, J. C. Northam, L. A. Tully, T. Carl, and A. Carter. “Developing Digital Mental Health Tools With Culturally Diverse Parents and Young People: Qualitative User-Centered Design Study”. In: *JMIR Pediatrics and Parenting*, *8*, 2025, e65163. doi: 10.2196/65163.

[18] A. J. Campbell and F. Robards. *Using technologies safely and effectively to promote young people’s wellbeing: a better practice guide for services*. Abbotsford, Vic.: Young and Well Cooperative Research Centre, 2013.

[19] S. Carlisle, K. Ayling, R. Jia, H. Buchanan, and K. Vedhara. “The effect of choice interventions on retention-related, behavioural and mood outcomes: a systematic review with meta-analysis”. In: *Health Psychology Review*, *16*, 2022, pp. 220–256. doi: 10.1080/17437199.2021.1962386.

[20] A. Castelletti, F. Pianosi, and M. Restelli. “A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run”. In: *Water Resources Research*, *49*, 2013, pp. 3476–3486. doi: 10.1002/wrcr.20295.

[21] D. Chapman and L. P. Kaelbling. “Input Generalization in Delayed Reinforcement Learning: An Algorithm and Performance Comparisons.” In: *Ijcai*. Vol. 91. 1991, pp. 726–731.

[22] P. L. Chapman and R. L. Mullis. “Racial Differences in Adolescent Coping and Self-Esteem”. In: *The Journal of Genetic Psychology*, *161*, 2000, pp. 152–160. doi: 10.1080/00221320009596702.

[23] R. A. Chechile. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods*. MIT Press, 2020.

[24] Z. Chen, M. Wu, A. Chan, X. Li, and Y.-S. Ong. “Survey on AI Sustainability: Emerging Trends on Learning Algorithms and Research Challenges [Review Article]”. In: *IEEE Computational Intelligence Magazine*, *18*, 2023, pp. 60–77. doi: 10.1109/MCI.2023.3245733.

[25] Z. M. Clayborne, M. Varin, and I. Colman. “Systematic Review and Meta-Analysis: Adolescent Depression and Long-Term Psychosocial Outcomes”. In: *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*, 2019, pp. 72–79. doi: 10.1016/j.jaac.2018.07.896.

[26] J. Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[27] C. S. Conley, E. B. Raposa, K. Bartolotta, S. E. Broner, M. Hareli, N. Forbes, K. M. Christensen, and M. Assink. “The Impact of Mobile Technology-Delivered Interventions on Youth Well-being: Systematic Review and 3-Level Meta-analysis”. In: *JMIR Mental Health*, *9*, 2022, e34254. doi: 10.2196/34254.

[28] J. K. Connor-Smith, B. E. Compas, M. E. Wadsworth, A. H. Thomsen, and H. Saltzman. “Responses to stress in adolescence: Measurement of coping and involuntary stress responses”. In: *Journal of Consulting and Clinical Psychology*, *68*, 2000, pp. 976–992. doi: 10.1037/0022-006X.68.6.976.

[29] E. P. Copeland and R. S. Hess. “Differences in Young Adolescents’ Coping Strategies Based On Gender and Ethnicity”. In: *The Journal of Early Adolescence*, *15*, 1995, pp. 203–219. doi: 10.1177/0272431695015002002.

[30] P. Czyżżak and A. Jaszkiewicz. “Pareto simulated annealing—a metaheuristic technique for multiple-objective combinatorial optimization”. In: *Journal of Multi-Criteria Decision Analysis*, *7*, 1998, pp. 34–47. doi: 10.1002/(SICI)1099-1360(199801)7:1<34::AID-MCDA161>3.0.CO;2-6.

[31] A. d’Halluin, M. Costa, M. Morgiève, and D. Sebbane. “Attitudes of Children, Adolescents, and Their Parents Toward Digital Health Interventions: Scoping Review”. In: *Journal of Medical Internet Research*, *25*, 2023, e43102. doi: 10.2196/43102.

[32] A. Delmas, B. Clement, P.-Y. Oudeyer, and H. Sauzéon. “Fostering Health Education With a Serious Game in Children With Asthma: Pilot Studies for Assessing Learning Efficacy and Automatized Learning Personalization”. In: *Frontiers in Education*, *3*, 2018. doi: 10.3389/feduc.2018.00099.

[33] L. Dennison, L. Morrison, G. Conway, and L. Yardley. “Opportunities and Challenges for Smartphone Applications in Supporting Health Behavior change: Qualitative Study”. In: *Journal of Medical Internet Research*, *15*, 2013, e86. doi: 10.2196/jmir.2583.

[34] M. Dierikx, N. Albers, B. L. Scheltinga, and W.-P. Brinkman. “Collaboratively Setting Daily Step Goals with a Virtual Coach: Using Reinforcement Learning to Personalize Initial Proposals”. In: *Persuasive Technology*. Ed. by N. Baghaei, R. Ali, K. Win, and K. Oyibo. Cham: Springer Nature Switzerland, 2024, pp. 100–115. doi: 10.1007/978-3-031-58226-4_9.

[35] E. Dietvorst, M. A. Aukes, J. S. Legerstee, A. Vreeker, M. M. Hrehovcsik, L. Keijsers, and M. H. J. Hillegers. “A Smartphone Serious Game for Adolescents (Grow It! App): Development, Feasibility, and Acceptance Study”. In: *JMIR Formative Research*, *6*, 2022, e29832. doi: 10.2196/29832.

[36] E. Dietvorst, L. P. de Vries, S. van Eijl, E. Mesman, J. S. Legerstee, L. Keijsers, M. H. J. Hillegers, and A. Vreeker. “Effective elements of eHealth interventions for mental health and well-being in children and adolescents: A systematic review”. In: *Digital Health*, *10*, 2024, p. 20552076241294105. doi: 10.1177/20552076241294105.

[37] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. “Explainable AI (XAI): Core Ideas, Techniques, and Solutions”. In: *ACM Comput. Surv.*, *55*, 2023, 194:1–194:33. doi: 10.1145/3561048.

[38] M. Eisenstadt, S. Liverpool, E. Infanti, R. M. Ciuvat, and C. Carlsson. “Mobile Apps That Promote Emotion Regulation, Positive Mental Health, and Well-being in the General Population: Systematic Review and Meta-analysis”. In: *JMIR Mental Health*, *8*, 2021, e31170. doi: 10.2196/31170.

[39] H. Eschenbeck, C.-W. Kohlmann, and A. Lohaus. “Gender Differences in Coping Strategies in Children and Adolescents”. In: *Journal of Individual Differences*, *28*, 2007, pp. 18–26. doi: 10.1027/1614-0001.28.1.18.

[40] M. Esmaeilimotlagh, K. Oveisi, F. Alizadeh, and M. Asadollahi Kheirabadi. “An Investigation on Coping Skills Training Effects on Mental Health Status of University Students”. In: *Journal of Humanities Insights*, *02*, 2018. doi: 10.22034/jhi.2018.61430.

[41] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 (General Data Protection Regulation)*. https://eur-lex.europa.eu/eli/reg/2016/679/oj. 2016.

[42] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. 2024.

[43] Z. Fatima Ezzahra, A. Sana, Q. Sara, and R. Said. “Multi-objective reinforcement learning for recommender systems: a comprehensive survey of methods, challenges, and future directions”. In: *International Journal of Multimedia Information Retrieval*, *14*, 2025, p. 33. doi: 10.1007/s13735-025-00383-7.

[44] F. Felten, L. N. Alegre, A. Nowé, A. L. C. Bazzan, E. G. Talbi, G. Danoy, and B. C. d. Silva. “A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning”. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. 2023.

[45] F. Felten, E.-G. Talbi, and G. Danoy. “Multi-Objective Reinforcement Learning Based on Decomposition: A Taxonomy and Framework”. In: *Journal of Artificial Intelligence Research*, *79*, 2024, pp. 679–723. doi: 10.1613/jair.1.15702.

[46] L. Fink, L. Newman, and U. Haran. “Let me decide: Increasing user autonomy increases recommendation acceptance”. In: *Computers in Human Behavior*, *156*, 2024, p. 108244. doi: 10.1016/j.chb.2024.108244.

[47] E. Frydenberg, R. Lewis, K. Bugalski, A. Cotta, C. McCarthy, N. Luscombe□smith, and C. Poole. “Prevention is better than cure: coping skills training for adolescents at school”. In: *Educational Psychology in Practice*, *20*, 2004, pp. 117–134. doi: 10.1080/02667360410001691053.

[48] D. C. Funder and D. J. Ozer. “Evaluating Effect Size in Psychological Research: Sense and Nonsense”. In: *Advances in Methods and Practices in Psychological Science*, *2*, 2019, pp. 156–168. doi: 10.1177/2515245919847202.

[49] S. Garrido, C. Millington, D. Cheers, K. Boydell, E. Schubert, T. Meade, and Q. V. Nguyen. “What Works and What Doesn’t Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People”. In: *Frontiers in Psychiatry*, *10*, 2019, p. 759. doi: 10.3389/fpsyt.2019.00759.

[50] C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao, and W. Liu. “A survey on interpretable reinforcement learning”. In: *Machine Learning*, *113*, 2024, pp. 5847–5890. doi: 10.1007/s10994-024-06543-w.

[51] G. Goldenits, T. Neubauer, S. Raubitzek, K. Mallinger, and E. Weippl. “Tabular reinforcement learning for reward robust, explainable crop rotation policies matching deep reinforcement learning performance”. In: *Computers and Electronics in Agriculture*, *237*, 2025, p. 110571. doi: 10.1016/j.compag.2025.110571.

[52] P. J. Gorman, P. J. Batterham, A. L. Calear, B. O’Dea, M. E. Larsen, D. J. Kavanagh, N. Titov, S. March, I. Hickie, M. Teesson, B. F. Dear, J. Reynolds, J. Lowinger, and L. Thornton. “Stakeholder perspectives on evidence for digital mental health interventions: Implications for accreditation systems”. In: *DIGITAL HEALTH*, 2019.

[53] C. Götzl, S. Hiller, C. Rauschenberg, A. Schick, J. Fechtelpeter, U. Fischer Abaigar, G. Koppe, D. Durstewitz, U. Reininghaus, and S. Krumm. “Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users’ and stakeholders’ perspectives”. In: *Child and Adolescent Psychiatry and Mental Health*, *16*, 2022, p. 86. doi: 10.1186/s13034-022-00522-6.

[54] R. Grist, J. Porter, and P. Stallard. “Mental Health Mobile Apps for Preadolescents and Adolescents: A Systematic Review”. In: *Journal of Medical Internet Research*, *19*, 2017, e176. doi: 10.2196/jmir.7332.

[55] E. C. S. de Groot, U. Gadiraju, O. Kudina, L. Keijsers, M. H. J. Hillegers, and W.-P. Brinkman. “Responsible Data Selection Method for Algorithmic Personalization of Health Apps: A Case Study on Promoting Mental Health”. In: *Frontiers in Digital Health*, *8*, 2026. doi: 10.3389/fdgth.2026.1691697.

[56] E. C. S. de Groot, U. Gadiraju, O. Kudina, A. Vreeker, H. Scholten, M. Hillegers, and W.-P. Brinkman. *Collection of code & data related to the CHALLENGE YOURSELF study*. 4TU Research Database. 2026. doi: 10.4121/6c139c04-13fe-4903-b2b7-091807222944.

[57] E. C. S. de Groot, L. P. de Vries, U. Gadiraju, O. Kudina, L. Keijsers, M. H. J. Hillegers, and W.-P. Brinkman. “Supporting adolescents’ mHealth needs: Qualitative and quantitative insights from a user survey of a mental health promoting app”. In: *Computers in Human Behavior Reports*, *22*, 2026, p. 101109. doi: 10.1016/j.chbr.2026.101109.

[58] A. el Hassouni, M. Hoogendoorn, M. Ciharova, A. Kleiboer, K. Amarti, V. Muhonen, H. Riper, and A. E. Eiben. “pH-RL: A Personalization Architecture to Bring Reinforcement Learning to Health Practice”. In: *Machine Learning, Optimization, and Data Science*. Ed. by G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, G. Jansen, P. M. Pardalos, G. Giuffrida, and R. Umeton. Cham: Springer International Publishing, 2022, pp. 265–280. doi: 10.1007/978-3-030-95467-3_20.

[59] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers. “A practical guide to multi-objective reinforcement learning and planning”. In: *Autonomous Agents and Multi-Agent Systems*, *36*, 2022, p. 26. doi: 10.1007/s10458-022-09552-y.

[60] N. Hazrati and F. Ricci. “Recommender systems effect on the evolution of users’ choices distribution”. In: *Information Processing & Management*, *59*, 2022, p. 102766. doi: 10.1016/j.ipm.2021.102766.

[61] N. Hazrati and F. Ricci. “Choice models and recommender systems effects on users’ choices”. In: *User Modeling and User-Adapted Interaction*, *34*, 2024, pp. 109–145. doi: 10.1007/s11257-023-09366-x.

[62] A. Jalalimanesh, H. S. Haghighi, A. Ahmadi, H. Hejazian, and M. Soltani. “Multi-objective optimization of radiotherapy: distributed Q-learning and agent-based simulation”. In: *Journal of Experimental & Theoretical Artificial Intelligence*, *29*, 2017, pp. 1071–1086. doi: 10.1080/0952813X.2017.1292319.

[63] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. “Planning and acting in partially observable stochastic domains”. In: *Artificial Intelligence*, *101*, 1998, pp. 99–134. doi: 10.1016/S0004-3702(98)00023-X.

[64] D. Keenan-Miller, C. L. Hammen, and P. A. Brennan. “Health Outcomes Related to Early Adolescent Depression”. In: *Journal of Adolescent Health*, *41*, 2007, pp. 256–262. doi: 10.1016/j.jadohealth.2007.03.015.

[65] R. Kenny, B. Dooley, and A. Fitzgerald. “Developing mental health mobile apps: Exploring adolescents’ perspectives”. In: *Health Informatics Journal*, *22*, 2016, pp. 265–275. doi: 10.1177/1460458214555041.

[66] S. Krishnan, M. Lam, S. Chitlangia, Z. Wan, G. Barth-Maron, A. Faust, and V. J. Reddi. *QuaRL: Quantization for Fast and Environmentally Sustainable Reinforcement Learning*. 2022. doi: 10.48550/arXiv.1910.01055.

[67] J. K. Kruschke. “Bayesian estimation supersedes the t test”. In: *Journal of Experimental Psychology: General*, *142*, 2013, pp. 573–603. doi: 10.1037/a0029146.

[68] T. Laure. “Inside ROOM: Unlocking the Design, Effects, and User Experience with a Mental Health App for University Students”. PhD thesis. Rotterdam: Erasmus University Rotterdam, 2025.

[69] T. Laure, C. Villegas Meija, D. Remmerswaal, D. Dulloo, R. Van der Hallen, B. Mayer, M. Littel, B. Dierckx, J. S. Legerstee, R. C. M. E. Engels, and M. Boffo. “ROOM to Grow, a Mobile Well-Being Intervention for University Students: Overview of the Design Process and Outcomes”. In: *JMIR Formative Research*, *9*, 2025, e63325. doi: 10.2196/63325.

[70] A. Lazaric. “Transfer in Reinforcement Learning: A Framework and a Survey”. In: *Reinforcement Learning: State-of-the-Art*. Ed. by M. Wiering and M. van Otterlo. Berlin, Heidelberg: Springer, 2012, pp. 143–173. doi: 10.1007/978-3-642-27645-3_5.

[71] J. S. Legerstee, N. Garnefski, F. C. Jellesma, F. C. Verhulst, and E. M. W. J. Utens. “Cognitive coping and childhood anxiety disorders”. In: *European Child & Adolescent Psychiatry*, *19*, 2010, pp. 143–150. doi: 10.1007/s00787-009-0051-6.

[72] S. Lehtimaki, J. Martic, B. Wahl, K. T. Foster, and N. Schwalbe. “Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview”. In: *JMIR Mental Health*, *8*, 2021, e25847. doi: 10.2196/25847.

[73] S. Li, E. C. S. de Groot, and W.-P. Brinkman. *Addressing Needs and Concerns in Coping Challenges Personalization using Multi-Objective Reinforcement Learning*. OSF Preprints. 2026.

[74] S. Li, E. C. S. de Groot, and W.-P. Brinkman. *Implementation and Analysis Code for Master’s Thesis: Balancing Stakeholder Needs in Adolescent mHealth Personalization using Multi-Objective Reinforcement Learning*. 4TU Research Database. 2026. doi: 10.4121/4a836cd2-0110-479d-be03-5b44437e6e91.

[75] Y. Lin, F. Lin, G. Cai, H. Chen, L. Zou, Y. Liu, and P. Wu. “Evolutionary Reinforcement Learning: A Systematic Review and Future Directions”. In: *Mathematics*, *13*, 2025, p. 833. doi: 10.3390/math13050833.

[76] J. M. Lipschitz, C. K. Pike, T. P. Hogan, S. A. Murphy, and K. E. Burdick. “The engagement problem: A review of engagement with digital mental health interventions and recommendations for a path forward”. In: *Current treatment options in psychiatry*, *10*, 2023, pp. 119–135. doi: 10.1007/s40501-023-00297-3.

[77] S. Litvin, R. Saunders, P. Jefferies, H. Seely, P. Pössel, and S. Lüttke. “The Impact of a Gamified Mobile Mental Health App (eQuoo) on Resilience and Mental Health in a Student Population: Large-Scale Randomized Controlled Trial”. In: *JMIR Mental Health*, *10*, 2023. doi: 10.2196/47285.

[78] S. Liverpool, C. P. Mota, C. M. D. Sales, A. Čuš, S. Carletto, C. Hancheva, S. Sousa, S. C. Cerón, P. Moreno-Peral, G. Pietrabissa, B. Moltrecht, R. Ulberg, N. Ferreira, and J. Edbrooke-Childs. “Engaging Children and Young People in Digital Mental Health Interventions: Systematic Review of Modes of Delivery, Facilitators, and Barriers”. In: *Journal of Medical Internet Research*, *22*, 2020, e16317. doi: 10.2196/16317.

[79] D. Lizotte, M. Bowling, and S. Murphy. “Linear Fitted-Q Iteration with Multiple Reward Functions”. In: *Proceedings of the International Conference on Automated Planning and Scheduling*, *23*, 2013, pp. 474–475. doi: 10.1609/icaps.v23i1.13579.

[80] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. “Explainable Reinforcement Learning through a Causal Lens”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 2020, pp. 2493–2500. doi: 10.1609/aaai.v34i03.5631.

[81] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. New York: Cambridge University Press, 2008.

[82] N. Marangunić and A. Granić. “Technology acceptance model: a literature review from 1986 to 2013”. In: *Universal Access in the Information Society*, *14*, 2015, pp. 81–95. doi: 10.1007/s10209-014-0348-1.

[83] G. Mayer, N. Gronewold, S. Alvarez, B. Bruns, T. Hilbel, and J.-H. Schultz. “Acceptance and Expectations of Medical Experts, Students, and Patients Toward Electronic Mental Health Apps: Cross-Sectional Quantitative and Qualitative Survey Study”. In: *JMIR Mental Health*, *6*, 2019, e14018. doi: 10.2196/14018.

[84] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys*, *54*, 2022, pp. 1–35. doi: 10.1145/3457607.

[85] S. Michie, L. Atkins, and R. West. *The behaviour change wheel: a guide to designing interventions*. Vol. 26. 2014.

[86] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker. *Model-based Reinforcement Learning: A Survey*. 2022. doi: 10.48550/arXiv.2006.16712.

[87] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. “A manifesto for reproducible science”. In: *Nature Human Behaviour*, *1*, 2017, p. 0021. doi: 10.1038/s41562-016-0021.

[88] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy. “Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support”. In: *Annals of Behavioral Medicine*, 2016. doi: 10.1007/s12160-016-9830-8.

[89] M. M. Ng, J. Firth, M. Minen, and J. Torous. “User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity”. In: *Psychiatric Services*, *70*, 2019, pp. 538–544. doi: 10.1176/appi.ps.201800519.

[90] A. Oláh. “Coping strategies among adolescents: a cross-cultural study”. In: *Journal of Adolescence*, *18*, 1995, pp. 491–512. doi: 10.1006/jado.1995.1035.

[91] S. Ontanon and J. Zhu. “The Personalization Paradox: the Conflict between Accurate User Models and Personalized Adaptive Systems”. In: *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces*. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 64–66. doi: 10.1145/3397482.3450734.

[92] O. Oyebode, F. Alqahtani, and R. Orji. “Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews”. In: *IEEE Access*, *8*, 2020, pp. 111141–111158. doi: 10.1109/ACCESS.2020.3002176.

[93] S. Padakandla, P. K. J., and S. Bhatnagar. “Reinforcement learning algorithm for non-stationary environments”. In: *Applied Intelligence*, *50*, 2020, pp. 3590–3606. doi: 10.1007/s10489-020-01758-5.

[94] S. Parisi, A. Blank, T. Viernickel, and J. Peters. “Local-utopia policy selection for multi-objective reinforcement learning”. In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2016, pp. 1–7. doi: 10.1109/SSCI.2016.7849369.

[95] M. Persike and I. Seiffge-Krenke. “Competence in Coping with Stress in Adolescents from Three Regions of the World”. In: *Journal of Youth and Adolescence*, *41*, 2012, pp. 863–879. doi: 10.1007/s10964-011-9719-6.

[96] B. Piko. “Gender Differences and Similarities in Adolescents’ Ways of Coping”. In: *The Psychological Record*, *51*, 2001, pp. 223–235. doi: 10.1007/BF03395396.

[97] J. Pollard, T. Reardon, C. Williams, C. Creswell, T. Ford, A. Gray, N. Roberts, P. Stallard, O. C. Ukoumunne, and M. Violato. “The multifaceted consequences and economic costs of child anxiety problems: A systematic review and meta-analysis”. In: *JCPP Advances*, *3*, 2023, e12149. doi: 10.1002/jcv2.12149.

[98] A. Radovic, A. Kirk-Johnson, M. Coren, B. George-Milford, and D. Kolko. “Stakeholder perspectives on digital behavioral health applications targeting adolescent depression and suicidality: Policymaker, provider, and community insights”. In: *Implementation Research and Practice*, *3*, 2022, p. 26334895221120796. doi: 10.1177/26334895221120796.

[99] R. Ribanszki, J. A. Saez Fonseca, J. M. Barnby, K. Jano, F. Osmani, S. Almasi, and E. Tsakanikos. “Preferences for Digital Smartphone Mental Health Apps Among Adolescents: Qualitative Interview Study”. In: *JMIR Formative Research*, *5*, 2021, e14004. doi: 10.2196/14004.

[100] D. A. Rohani, A. Quemada Lopategui, N. Tuxen, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. “MUBS: A Personalized Recommender System for Behavioral Activation in Mental Health”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. doi: 10.1145/3313831.3376879.

[101] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. “A Survey of Multi-Objective Sequential Decision-Making”. In: *Journal of Artificial Intelligence Research*, *48*, 2013, pp. 67–113. doi: 10.1613/jair.3987.

[102] R. M. Ryan and E. L. Deci. “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.” In: *American psychologist*, *55*, 2000, p. 68.

[103] R. M. Ryan and E. L. Deci. “Self-determination theory”. In: *Encyclopedia of quality of life and well-being research*. Springer, 2024, pp. 6229–6235.

[104] R. M. Ryan, H. Patrick, E. L. Deci, and G. C. Williams. “Facilitating health behaviour change and its maintenance: Interventions based on self-determination theory”. In: *European Health Psychologist*, *10*, 2008, pp. 2–5.

[105] S. Saxena, G. Thornicroft, M. Knapp, and H. Whiteford. “Resources for mental health: scarcity, inequity, and inefficiency”. In: *The Lancet*, *370*, 2007, pp. 878–889. doi: 10.1016/S0140-6736(07)61239-2.

[106] B. Scheibehenne, R. Greifeneder, and P. M. Todd. “Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload”. In: *Journal of Consumer Research*, *37*, 2010, pp. 409–425. doi: 10.1086/651235.

[107] K. Shang, H. Ishibuchi, L. He, and L. M. Pang. “A Survey on the Hypervolume Indicator in Evolutionary Multiobjective Optimization”. In: *IEEE Transactions on Evolutionary Computation*, *25*, 2021, pp. 1–20. doi: 10.1109/TEVC.2020.3013290.

[108] X. Shao, L. Liu, X. Zhu, C. Tian, D. Li, L. Zhang, X. Liu, Y. Liu, G. Zhu, and L. Li. “Personalized game-based digital intervention for relieving depression and anxiety symptoms: a pilot RCT”. In: *npj Mental Health Research*, *4*, 2025, p. 27. doi: 10.1038/s44184-025-00141-x.

[109] B. Smith, A. Khojandi, and R. Vasudevan. “Bias in Reinforcement Learning: A Review in Healthcare Applications”. In: *ACM Comput. Surv.*, *56*, 2023, 52:1–52:17. doi: 10.1145/3609502.

[110] A. Straw, T. Wiecki, C. Fonnesbeck, A. Suárez, and O. Martin. “Bayesian Estimation Supersedes the T-Test”. In: *PyMC examples*. Ed. by P. Team. doi: 10.5281/zenodo.5654871.

[111] R. S. Sutton, A. G. Barto, et al. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge, 1998.

[112] H. Taherdoost. “A review of technology acceptance and adoption models and theories”. In: *Procedia Manufacturing*, *22*, 2018, pp. 960–967. doi: 10.1016/j.promfg.2018.03.137.

[113] M. E. Taylor and P. Stone. “Transfer Learning for Reinforcement Learning Domains: A Survey”. In: *Journal of Machine Learning Research*, *10*, 2009.

[114] G. Terlouw, J. T. B. van’t Veer, D. A. Kuipers, and J. Metselaar. “Context analysis, needs assessment and persona development: towards a digital game-like intervention for high functioning children with ASD to train social skills”. In: *Early Child Development and Care*, *190*, 2020, pp. 2050–2065. doi: 10.1080/03004430.2018.1555826.

[115] S. Tomkins, P. Liao, P. Klasnja, and S. Murphy. “IntelligentPooling: practical Thompson sampling for mHealth”. In: *Machine Learning*, *110*, 2021, pp. 2685–2727. doi: 10.1007/s10994-021-05995-8.

[116] J. Torous, J. Nicholas, M. E. Larsen, J. Firth, and H. Christensen. “Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements”. In: *Evidence Based Mental Health*, *21*, 2018. doi: 10.1136/eb-2018-102891.

[117] Y. Wei, P. Zheng, H. Deng, X. Wang, X. Li, and H. Fu. “Design Features for Improving Mobile Health Intervention User Engagement: Systematic Review and Thematic Analysis”. In: *Journal of Medical Internet Research*, *22*, 2020, e21687. doi: 10.2196/21687.

[118] T. Weimann and C. Gißke. “Unleashing the Potential of Reinforcement Learning for Personalizing Behavioral Transformations with Digital Therapeutics: A Systematic Literature Review:” in: *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies*. Rome, Italy: SCITEPRESS - Science and Technology Publications, 2024, pp. 230–245. doi: 10.5220/0012474700003657.

[119] B. Wies, C. Landers, and M. Ienca. “Digital Mental Health for Young People: A Scoping Review of Ethical Promises and Challenges”. In: *Frontiers in Digital Health*, *3*, 2021. doi: 10.3389/fdgth.2021.697072.

[120] World Health Organization. *Mental health of adolescents*. 2025. url: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health (visited on 12/03/2025).

[121] A. van Wynsberghe. “Sustainable AI: AI for sustainability and the sustainability of AI”. In: *AI and Ethics*, *1*, 2021, pp. 213–218. doi: 10.1007/s43681-021-00043-6.

[122] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik. “Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control”. In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, pp. 10607–10616.

[123] L. Yardley, B. J. Spring, H. Riper, L. G. Morrison, D. H. Crane, K. Curtis, G. C. Merchant, F. Naughton, and A. Blandford. “Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions”. In: *American Journal of Preventive Medicine*, *51*, 2016, pp. 833–842. doi: 10.1016/j.amepre.2016.06.015.

[124] F. E. Zaizi, S. Qassimi, and S. Rakrak. “Multi-objective optimization with recommender systems: A systematic review”. In: *Information Systems*, *117*, 2023, p. 102233. doi: 10.1016/j.is.2023.102233.

[125] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao. “A survey on federated learning”. In: *Knowledge-Based Systems*, *216*, 2021, p. 106775. doi: 10.1016/j.knosys.2021.106775.

[126] Y. Zheng and D. Wang. “A survey of recommender systems with multi-objective optimization”. In: *Neurocomputing*, *474*, 2022, pp. 141–153. doi: 10.1016/j.neucom.2021.11.041.

[127] S. Zhu, Y. Wang, and Y. Hu. “Facilitators and Barriers to Digital Mental Health Interventions for Depression, Anxiety, and Stress in Adolescents and Young Adults: Scoping Review”. In: *Journal of Medical Internet Research*, *27*, 2025, e62870. doi: 10.2196/62870.

[128] L. M. Zintgraf, T. V. Kanters, D. M. Roijers, F. A. Oliehoek, and P. Beau. “Quality Assessment of MORL Algorithms: A Utility-Based Approach”. In: *Proceedings of the 24th Annual Machine Learning Conference of Belgium and the Netherlands.*, 2015.

[129] E. Zitzler, L. Thiele, M. Laumanns, C. Fonseca, and V. da Fonseca. “Performance assessment of multiobjective optimizers: an analysis and review”. In: *IEEE Transactions on Evolutionary Computation*, *7*, 2003, pp. 117–132. doi: 10.1109/TEVC.2003.810758.

[130] E. Zitzler and L. Thiele. “Multiobjective optimization using evolutionary algorithms — A comparative case study”. In: *Parallel Problem Solving from Nature — PPSN V*. Ed. by A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel. Berlin, Heidelberg: Springer, 1998, pp. 292–301. doi: 10.1007/BFb0056872.

A

Challenge measures

Table A.1: Challenge measures with the question posed to the participants and the measurement scale.

Measure	**Question**	**Scale**
Before receiving a challenge		
Tiredness	How tired are you right now?	7-point Likert scale from “Not at all” to “Very much”
Motivation	How excited are you to do a challenge?	7-point Likert scale from “Not at all” to “Very much”
Usefulness belief	How good do you think it is to do a challenge?	7-point Likert scale from “Not at all” to “Very much”
Time	How much time do you have right now?	7-point Likert scale from “None at all” to “A lot”
Return willingness	Currently, you are taking part in a paid study. Imagine you were not being paid for opening your daily challenge. How likely would you then have been to open your daily challenge?	7-point Likert scale from “Absolutely not” to “Absolutely”
Challenge evaluation		
Perceived enjoyment	How much did you like the challenge?	Scale from 0 to 10
Difficulty	How difficult did you find the challenge?	Scale from 0 to 10
Time spent	How long did you take to complete the challenge?	Minutes
Perceived usefulness	How useful was this challenge for you?	7-point Likert scale from “Not at all” to “Very much”

B

Number of samples

Table B.1: Number of samples per state-action pair and per state. States are denoted as (tiredness, time available, motivation), where tiredness and time are binary (0 = low, 1 = high) and motivation is tertiary (0 = low, 1 = medium, 2 = high). Actions are denoted as (a_c, a_s) with challenge cluster $a_c \in \{D, EU, EUD\}$ and within-cluster selection strategy $a_s \in \{0, 1\}$.

	Action						
State	(D, 0)	(EU, 0)	(EUD, 0)	(D, 1)	(EU, 1)	(EUD, 1)	Total
(0, 0, 0)	57	90	99	70	147	146	609
(0, 0, 1)	39	70	78	73	114	121	495
(0, 0, 2)	24	31	41	47	59	85	287
(0, 1, 0)	17	19	19	9	30	40	134
(0, 1, 1)	26	26	29	22	47	49	199
(0, 1, 2)	43	42	53	54	107	98	397
(1, 0, 0)	67	76	86	66	131	113	539
(1, 0, 1)	38	38	33	46	86	65	306
(1, 0, 2)	12	17	11	25	25	40	130
(1, 1, 0)	11	11	15	9	26	27	99
(1, 1, 1)	7	12	17	15	28	35	114
(1, 1, 2)	15	17	20	23	42	45	162

C

Reward correlations

Table C.1: Reward correlations with challenge completion, alongside the resulting weight for the correlation-based baseline policy.

Reward	Correlation	Weight
Perceived enjoyment	0.833	0.305
Perceived usefulness	0.801	0.294
Return willingness	0.098	0.036
Adherence	1.000	0.367
Diversity	-0.005	-0.002

D

Additional results

This appendix provides additional results for the policies that only optimize a single objective (section D.1) and the outcome trajectories over time (section D.2).

D.1. Single-objective policies

D.1.1. Obtained policies

Table D.1 shows the state-action mappings of the policies that are learned when optimizing for only one of the objectives.

Table D.1: Policies obtained when optimizing for only one of the objectives, mapping user states to actions. States are denoted as (tiredness, time available, motivation), where tiredness and time are binary (0 = low, 1 = high) and motivation is tertiary (0 = low, 1 = medium, 2 = high). Actions are denoted as (a_c, a_s) with challenge cluster $a_c \in \{D, EU, EUD\}$ and within-cluster selection strategy $a_s \in \{0, 1\}$.

	Optimized objective				
State	Perceived Enjoyment	Perceived Usefulness	Diversity	Adherence	Return Willingness
(0, 0, 0)	(EU,0)	(EU,0)	(EU,1)	(EU,0)	(D,1)
(0, 0, 1)	(EU,1)	(EUD,0)	(EU,0)	(EU,0)	(D,1)
(0, 0, 2)	(EUD,0)	(EUD,0)	(EU,0)	(EUD,1)	(D,1)
(0, 1, 0)	(EU,0)	(EU,0)	(EUD,1)	(EU,0)	(EUD,1)
(0, 1, 1)	(EU,0)	(EU,0)	(EU,0)	(EU,0)	(EU,1)
(0, 1, 2)	(EUD,1)	(EUD,1)	(EUD,0)	(EUD,1)	(EU,1)
(1, 0, 0)	(EU,0)	(EU,0)	(EU,0)	(EU,0)	(D,1)
(1, 0, 1)	(EU,1)	(EU,1)	(D,0)	(EU,1)	(EU,1)
(1, 0, 2)	(EU,0)	(EU,0)	(D,0)	(EU,1)	(EUD,1)
(1, 1, 0)	(EU,1)	(EUD,1)	(EUD,1)	(EU,1)	(EU,1)
(1, 1, 1)	(EUD,0)	(EUD,0)	(D,0)	(EU,0)	(D,1)
(1, 1, 2)	(EU,0)	(EU,0)	(EU,1)	(EU,0)	(D,1)

D.1.2. Diversity of coping categories in practice

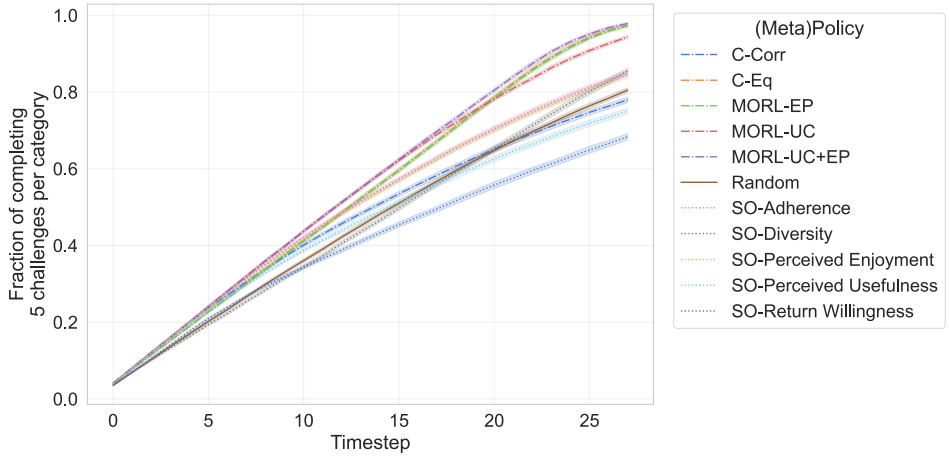


Figure D.1: Mean fraction (with their 95% CI) of completing at least 5 challenges per coping category over 28 timesteps for each (meta)policy and the single-objective (SO) policies. The lines for MORL-UC and MORL-UC+EP are completely overlapping. Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy.

D.1.3. User retention

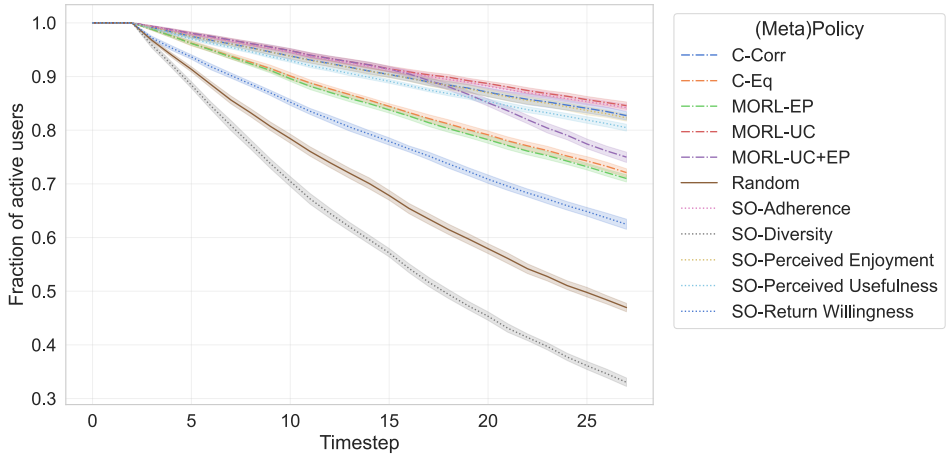
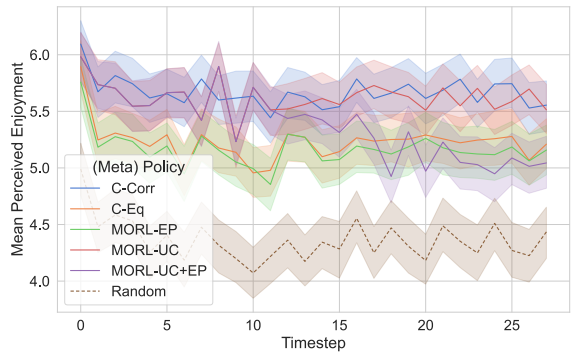
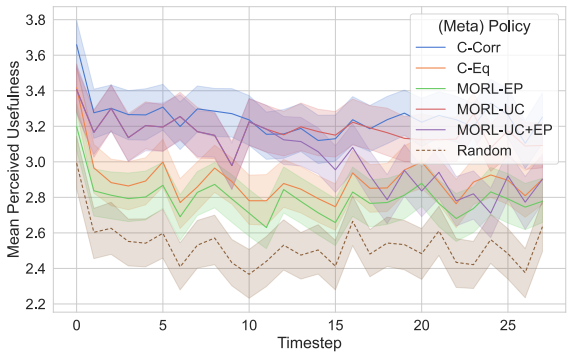


Figure D.2: Mean fraction of active users in the simulations for each (meta)policy and the single-objective policies, alongside the 95% confidence interval. Simulations were run with 500 users for 20 runs. Users are assumed to be inactive after 3 consecutive days of non-completion. Policy abbreviations: MORL-EP = Expert Priority Metapolicy; MORL-UC = User Choice Metapolicy; MORL-UC+EP = UC followed by EP after 12 challenge completions; C-Corr = Correlation-based policy; C-Eq = Equal weights policy..

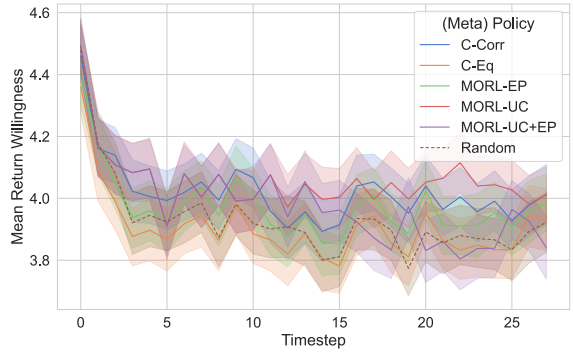
D.2. Outcomes over time



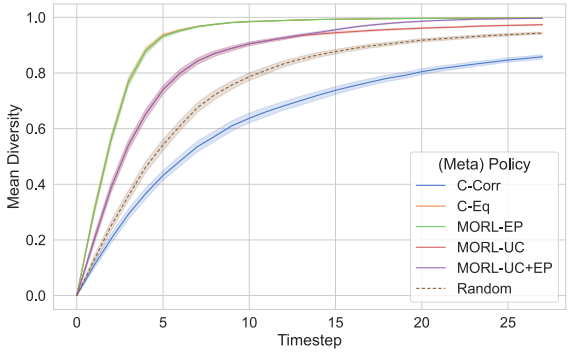
(a) Perceived Enjoyment



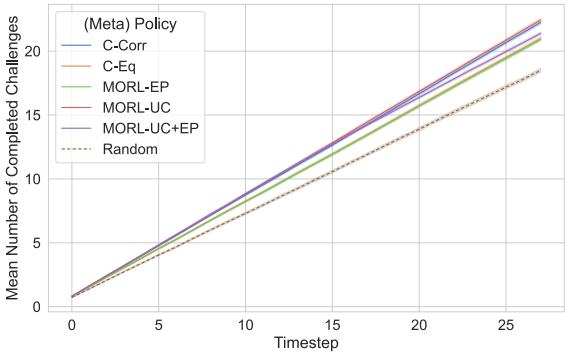
(b) Perceived usefulness



(c) Return willingness



(d) Diversity



(e) Number of completed challenges

Figure D.3: Outcome trajectories over 28 timesteps across 1000 simulations. The MORL-UC+EP metapolicy switches from user choice to expert priority after completing 12 challenges, which causes a decrease in perceived enjoyment and perceived usefulness. Policy abbreviations: MORL-EP = MORL with Expert Priority; MORL-UC = User Choice; MORL-RC+EP = Combination; B-Corr = Correlation-based baseline; B-Eq = Equal weights baseline

E

Additional Bayesian analyses

This appendix provides additional results of Bayesian analyses of the improvement in adherence for agency versus non-agency conditions in our dataset, and improvements of our state-action model over other models for transition and reward dynamics.

E.1. Adherence in agency vs. non-agency conditions

Table E.1: Improvement in adherence (Δ) of agency versus non-agency conditions, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

			P(**Effect size** $>$)		
Mean Δ **[95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
0.031 $[0.006, 0.055]$	0.990	Strong bet irresponsible to avoid	0.80	0.26	0.02

E.2. Transition function

E.2.1. State-action vs. uniform baseline

The following table provides results of the Bayesian analysis when inspecting the improvements in likelihoods of the next state when using the state-action model over the uniform baseline to predict the next state.

Table E.2: Mean improvements of likelihoods (Δ) when using the state-action transition model versus the uniform baseline for predicting the next state, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** $>$)		
State	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
(0, 0, 0)	0.131 $[0.112, 0.148]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 0, 1)	0.053 $[0.043, 0.065]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 0, 2)	0.032 $[0.023, 0.040]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 1, 0)	0.073 $[0.049, 0.099]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 1, 1)	0.023 $[0.010, 0.035]$	$> .99995$	Virtually certain	$> .99$	0.95	0.70
(0, 1, 2)	0.029 $[0.020, 0.040]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.97
(1, 0, 0)	0.080 $[0.066, 0.092]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 0, 1)	0.053 $[0.041, 0.064]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 0, 2)	0.019 $[0.006, 0.032]$	0.998	Very strong bet	0.97	0.83	0.53
(1, 1, 0)	0.030 $[0.010, 0.051]$	0.998	Very strong bet	0.98	0.89	0.66
(1, 1, 1)	0.009 $[-0.002, 0.020]$	0.934	A promising but risky bet	0.75	0.41	0.15
(1, 1, 2)	0.028 $[0.012, 0.044]$	$> .99995$	Nearing certainty	0.99	0.96	0.80

E.2.2. State-action vs. stay-in-state

The following table provides results of the Bayesian analysis when inspecting the improvements in likelihoods of the next state when using the state-action model over assuming users stay in their current state to predict the next state.

Table E.3: Mean improvements of likelihoods (Δ) when using the state-action model versus the stay-in-state model for predicting the next state, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** $>$)		
State	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
(0, 0, 0)	-0.069 $[-0.099, -0.039]$	$< .00005$	Virtually certain against	$> .99$	0.95	0.59
(0, 0, 1)	0.007 $[-0.052, 0.077]$	0.475	Not worth betting against	0.70	0.42	0.28
(0, 0, 2)	0.091 $[0.075, 0.106]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 1, 0)	0.116 $[0.079, 0.153]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 1, 1)	0.087 $[0.072, 0.101]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 1, 2)	0.070 $[0.059, 0.081]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 0, 0)	-0.057 $[-0.091, -0.025]$	0.0003	Nearing certainty against	0.98	0.77	0.27
(1, 0, 1)	0.090 $[0.062, 0.117]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.99
(1, 0, 2)	0.098 $[0.079, 0.118]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 1, 0)	0.096 $[0.066, 0.126]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 1, 1)	0.091 $[0.078, 0.103]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 1, 2)	0.063 $[0.034, 0.094]$	0.999	Very strong bet	0.99	0.98	0.94

E.2.3. State-action vs. action-only

The following table provides results of the Bayesian analysis when inspecting the improvements in likelihoods of the next state when using the state-action model over the action-only model to predict the next state.

Table E.4: Mean improvements of likelihoods (Δ) when using the state-action transition model versus the action-only model for predicting the next state, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** $>$)		
State	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
(0, 0, 0)	0.073 $[0.059, 0.085]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(0, 0, 1)	0.013 $[0.005, 0.021]$	$> .99995$	Nearing certainty	0.97	0.70	0.20
(0, 0, 2)	0.006 $[-0.002, 0.015]$	0.921	A promising but risky bet	0.59	0.18	0.02
(0, 1, 0)	0.025 $[0.006, 0.046]$	0.995	Very strong bet	0.95	0.80	0.49
(0, 1, 1)	0.001 $[-0.012, 0.013]$	0.542	Not worth betting on	0.29	0.03	$< .01$
(0, 1, 2)	0.010 $[-0.002, 0.022]$	0.946	A promising but risky bet	0.59	0.14	0.01
(1, 0, 0)	0.033 $[0.024, 0.041]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
(1, 0, 1)	0.015 $[0.006, 0.025]$	0.999	Very strong bet	0.97	0.73	0.28
(1, 0, 2)	0.002 $[-0.014, 0.017]$	0.588	Not worth betting on	0.36	0.06	$< .01$
(1, 1, 0)	0.008 $[-0.008, 0.023]$	0.834	Only a casual bet	0.59	0.25	0.07
(1, 1, 1)	-0.004 $[-0.020, 0.012]$	0.329	Not worth betting against	0.44	0.12	0.02
(1, 1, 2)	0.031 $[0.010, 0.049]$	0.999	Very strong bet	0.98	0.90	0.66

E.3. Deterministic rewards

Table E.5 shows the results of the Bayesian analysis when comparing the state-action and action-only model to the global model for estimating the deterministic rewards. This means we use the mean per state-action pair versus the global mean for each objective.

Table E.5: Mean reduction of L_1-errors (Δ) when using 1) the action-only and 2) the state-action model compared to the global model for predicting rewards, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** >)		
Objective	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
Action-only > global						
Perceived Enjoyment	0.010 $[0.008, 0.013]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Perceived Usefulness	0.010 $[0.008, 0.012]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Return Willingness	-0.002 $[-0.002, -0.001]$	$< .00005$	Virtually certain against	$> .99$	$> .99$	$> .99$
Adherence	0.009 $[0.006, 0.011]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.97
State-action > global						
Perceived Enjoyment	0.015 $[0.012, 0.019]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Perceived Usefulness	0.013 $[0.010, 0.016]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Return Willingness	0.003 $[-0.002, 0.009]$	0.828	Only a casual bet	0.22	0.01	0.00
Adherence	0.015 $[0.011, 0.019]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.93

E.4. Stochastic outcomes

Table E.6 shows the results of the Bayesian analysis when comparing the state-action and action-only model to the global model for estimating the stochastic outcome distributions used for our simulation environment.

Table E.6: Mean improvement Δ of outcome likelihoods when using the state-action model versus the global model to estimate the outcome probability distributions, alongside their 95% HDI, posterior probability, and interpretation following the guidelines by Chechile [23] and Andraszewicz et al. [8].

				P(**Effect size** >)		
Outcome	**Mean Δ [95% HDI]**	**post. prob**$(\Delta > 0)$	**Interpretation**	**Small**	**Medium**	**Large**
Action-only > global						
Perceived Enjoyment	0.003 $[0.002, 0.005]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.87
Perceived Usefulness	0.009 $[0.007, 0.011]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Return Willingness	0.001 $[0.000, 0.002]$	0.997	Very strong bet	0.85	0.25	0.01
Completion	0.009 $[0.006, 0.011]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.97
State-action > global						
Perceived Enjoyment	0.003 $[< .00005, 0.005]$	0.981	Good bet too good to disregard	0.64	0.10	$< .01$
Perceived Usefulness	0.011 $[0.008, 0.015]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.97
Return Willingness	0.024 $[0.020, 0.029]$	$> .99995$	Virtually certain	$> .99$	$> .99$	$> .99$
Completion	0.015 $[0.011, 0.019]$	$> .99995$	Virtually certain	$> .99$	$> .99$	0.94