Use Reinforcement Learning to Choose Activities for Preparing to Quit Smoking

How Effective a Reinforcement Learning Model is for Choosing Activities that Optimizes the Likelihood that Users Return to the Next Session and the Effort Users Spend on Their Activities?

Meng Zhang

Supervisor(s): Responsible Professor: Willem-Paul Brinkman, Supervisor: Nele Albers

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
February 2,2024

Name of the student: Meng Zhang
Final project course: CSE3000 Research Project
Thesis committee: Responsible Professor: Willem-Paul Brinkman, Supervisor: Nele Albers, Examiner: Huijuan Wang

Abstract

Unhealthy behaviors such as smoking is the major cause for premature deaths and changing behaviors by oneself can be difficult. That is where eHealth applications come into rescue. One of the recent research explored the possibility of using a Reinforcement Learning model to choose persuasive types for a virtual coach to adopt to persuade people to prepare for smoke-quitting and it has shown advantages. However, there are still more aspects to investigate in this context except the persuasive types of the messages, and this paper intended to further look into using reinforcement learning to choose activities for preparing the users to quit smoking, To be more specific, we implemented and evaluated a reinforcement learning model to choose activities to optimize both the effort spent by the users and also the likelihood of them staying for the next session. The result suggests that reinforcement learning is a promising approach to choose activities for people to prepare for quitting smoking and it can move the users to states that they are more likely to spend a good effort on the activities and are more likely to come back to the next session.

1 Introduction

Research has shown that 40% of premature deaths are caused by unhealthy behaviours [10], with smoking and inactivity being the two major reasons [2]. To be more specific, every year smoking leads to eight million deaths [16] and half of the users who do not quit will most likely die because of it [4] [7] [15]. Therefore, it is important for people to abandon those unhealthy behaviours, especially smoking, to live a healthier life and prevent premature deaths. However, changing behaviours and habits can be difficult by individuals alone [6]. In fact, 68% of adult smokers in the U.S. in 2015 indicated that they wanted to quit smoking but only 10.7% of them succeeded to quit smoking for at least 6 months [1].

Considering all these, external help is necessary for people who need or intend to give up unhealthy behaviours. Among all the means and approaches eHealth applications has shown its advances under this context with their persuasive technology [18] [8]. However, research has shown that a single persuasive attempt has limited effect on people's behavior, and according to behavior change theories, behaviors and people's states influence each other [11]. Therefore, to see if we can persuade people in a way that we can move them to a state where they are more likely to be persuaded again, a research, carried out by Albers et al, used reinforcement learning model to optimize the persuasive type of the messages sent to the users, who are daily smokers, to help them prepare to quit smoking and to be more physically active [11]. During this research, a longitudinal study of five sessions, where more than 500 daily smokers interacted with a virtual coach was conducted. And the research shows that using reinforcement learning can better predict both the effort spent by the users and the future states after the persuasive attempts and thus improve the effectiveness of the persuasive attempts [11].

But there are more to be done to understand the behavior of the users and the effectiveness of reinforcement learning for behavior changing. This paper aims to investigate another aspect of this topic and it is based on the research by Albers [13]. By this paper we intend to see if a reinforcement learning model can be effectively used to choose activities for the users to follow to optimize the likelihood of the users' return to the next session as well as the effort that the users actually spend on their activities during these sessions. This is different from the previous work in two aspects, one is that we choose from activities instead of persuasive types, since at the end we want the users to perform these activities to get them ready for quit smoking and different activities may have different influence on the users when the users are in different states. The other is that we consider not only the effort users spent but how much they are willing to stay and return to the next session, because we not only want the users to make an effort and perform the assigned activities but also want them to return to future sessions, instead of dropping out.

We implemented and evaluated a reinforcement learning model, which used the features of the users such as their usefulness beliefs, their energy level and their business level as the candidates for forming the state space, to choose activities that can optimize the weighted sum of the effort spend and the likelihood of users return to next session. The evaluation results confirmed that the users' states does influence the effectiveness of the activities and that using the reinforcement learning model to choose activities can indeed increase the effort the users spend on the activities and also increase their willingness to return for future sessions. However, there is also some limitations of this paper and one of them being that the model is built under the assumption that the reward function is the weighted sum of the dropout likelihood and the effort spent, instead of other assumptions such as the multiplication of these two goals. Therefore, there might be some future research necessary to further investigate the relations between these two goals.

2 Background

To answer the research question, how to use reinforcement learning (RL) to choose activities for preparing to quit smoking, to be more specific, how effective a reinforcement learning model is for choosing activities that optimizes the likelihood that users return to the next session and the effort users spend on their activities, we break this big and general question into five specific analysis questions. These specific analysis questions intend to evaluate the different components and aspects of the RL model. So in this chapter we will introduce the background of this research and motivate these five analysis questions. And in chapter 4 we illustrate the experiments for answering each question.

2.1 Activities

As mentioned in the introduction, a virtual coach, Mel, needed to choose one activity for each session per user. There are 53 different activities to be chosen from and they can be divided into two categories, 9 persuasive activities and 44 preparatory activities. The typical activity is to watch a provided video and answer some questions by writing it down. By doing these activities, the user is expected to be more prepared to quit smoking and to be prepared to become more physically active.

2.2 States

How these activities or the action affect the users may not be constant, instead it might depend on the states of the users [13]. For our research, we derive the states of the users based on features such as their energy level, their business level and their beliefs about the usefulness of nine competencies, including self-efficacy, practical knowledge, awareness of positive outcomes, awareness of negative outcomes, motivation to change, knowledge of how to maintain or achieve mental well-being, mindset that physical activity helps to quit smoking, awareness of smoking patterns and knowledge of how to maintain or achieve well-being. So the first analysis question we would like to investigate is:

Q1: How well can states derived from the features predict behavior, which is the effort spent by the users and their willingness to come back for the next session, after performing the chosen activities?

2.3 Future States

We intend to get the users to the states that the reward is better, they make good effort on the activities and also have high willingness to return to future sessions. And to do that we need to look at people's future state and for starters we need to be able to give good predictions about the next state of the users after they performed the chosen activities since different activities might influence the next states. Therefore the second question is formulated as the following:

Q2: How well can the states derived from the provided features about users predict states after choosing activities?

Next, we looked at the optimal policy for choosing activities. The optimal policy is the best action or activity that should be taken for each state. What we wanted to achieve is that this optimal policy can choose activities that move the users to a state that they are more likely to spend adequate effort on the activities and also to return to the next session. Therefore, we wanted to see if users can indeed arrive the state after some optimal actions being taken in the future. So we have the third analysis question:

Q3: What is the effect of multiple activities chosen by the optimal policy on the users' states?

After checking the effect of the activities on the user's states, we can evaluate if our optimal policy is indeed better than other policies. Here optimal policy gives the best action to take for each state that the users can possibly in. So we formulate our fourth question:

Q4: How effective is the optimal activities compared to non-optimal ones in their effect on behavior?

2.4 Reward Function

Finally, we would like to investigate our reward function. The reward function for our project represents two goals, to increase the likelihood of users' return to the next session and to increase the effort that users are willing to spend during these sessions. So there are more than one way to form the reward function. Thus we would like to see if different ways of forming the reward function can lead to different optimal activities and we form the analysis question as following:

Q5: How do different optimal policies based on different reward functions compare with each other regarding the effect of multiple optimal activities on behavior?

3 Methodology

Since this paper used the data collected in the study done by Albers et al [13], so in this chapter, we will briefly describe the virtual coach developed for the study and the study first and then we will explain our reinforcement learning model.

3.1 Virtual Coach

A text-based virtual coach Mel was implemented to help people prepare for quitting smoking by conversational sessions [3]. The study started with a pre-screening questionnaire, where participants were filtered with conditions such as speaking English fluently, are daily smokers and at least 18 years old. And then the eligible ones were invited to five sessions with Mel. Each session started with the participants being asked about their states, including their belief of usefulness of nine competencies such as self-efficacy, practical knowledge, awareness of positive outcomes, awareness of negative outcomes, motivation to change, knowledge of how to maintain mental well-being, mindset that physical activity helps to quit smoking, awareness of smoking patterns, and also their business level and energy level. And then it followed by the coach chose activities for the participants to perform. And then in the next session, the participants were asked about their experience with last session, including Likert-scale questions such as how much effort they spent for the last session's activity, how likely they would return to the next session if it is not a paid program, and open questions such as how they approached their assigned activity in the previous session and assign their, and also the states questions, and then again another activity was assigned to them. After that, participants who did all five sessions were also invited for a post-questionnaire and a follow-up questionnaire.

3.2 Data Collection

864 people who joined the pre-screening were daily smokers and at least 18 years old and speak fluent English, and 646 of them were eligible to go further for the study since they were contemplating or preparing to quit smoking. The gender distribution of the eligible participants was balanced, with

53% of them being female and 45% being male. The majority (71%) of the eligible participants were from the age range of 26 to 50, and with the rest 11% being young adults and 17% being old people. From the five sessions mentioned above we obtained 1721 transition samples from 547 people. Here a transition sample is a (s, a, r, s') tuple where s is the state, a is the action taken, r is the reward received after a, s' is the next state after a. And this is the data we needed for the research.

3.3 Reinforcement Learning Model

To answer those analytical research questions mentioned in chapter 2, we needed to first implement the reinforcement learning(RL) algorithm based on the data available to give us an optimal policy which choose activities that optimize the dropout rate and the effort spent by the user. RL works in a Markov Desicion Process (MDP), which is a 5 tuple (S, A, R, T, γ):

- S is the state space
- A is the actions that can be taken
- R is the reward function
- T is the transition function
- γ is the discount factor

In this section, we will describe elements, including states, actions, reward functions and discount factor, for our RL model and also how the data is processed to fit in the model for our purpose.

States

States of an MDP captures the relevant information of the world or the environment that the agent or the decision maker acts in. For our case, the states are the status of the user upon which the virtual coach would choose an activity and it is extracted from the questions such as the usefulness beliefs, the busyness level and the energy level as mentioned above in section 3.1.

There are eleven state features in total and every one of them is numeric. But there are only a very limited number of transition samples (1721) available, so we need to reduce the state space to preserve reliability. First step taken to is to convert the numeric answers to binary. Specifically, we compute the mean for each state feature and convert the corresponding answers to 1 if it is equal or greater than the mean or 0 if it is less than the mean. After this, we selected three features to further reduce the state space.

The feature selection is inspired by the G algorithm [5] and the features chosen are the usefulness beliefs of the following competencies for quit smoking:

- Mindset that physical activity helps to quit smoking
- Awareness of smoking patterns
- Self-efficacy

Therefore, there are eight states in the state space. Each one of them is a three character long binary code with each binary character representing the usefulness beliefs as above in order. For example, state [101] means that the users in this state have above average beliefs that the mindset that physical activity helps to quit smoking and self-efficacy are useful, but they have less than average belief that being aware of their smoking patterns is useful.

Actions

Actions of an MDP is the set of actions that the agent can take in each state of the state space. For our case it corresponds to the 14 different activity clusters that the virtual coach can choose an activity from to give the user. As explained in chapter 2, activities can be seen as either persuasive ones, or preparatory ones. There are nine persuasive activities with each of them as one cluster and corresponding to the nine competencies mentioned in section 3.1. And there are also preparatory activities which are classified into five clusters based on their similarities perceived by smokers [12].

Reward Functions

Reward function for our case is the two goals we are trying to optimize by our choosing process and they are to the dropout response and the effort spent response in each session. So a natural question to ask is what the relation between these two factors and whether they are correlated. The Spearman test resulted in 0.39 and suggested a somewhat low positive correlation between these two goals. However, the Cohen's kappa between these two goals is -0.04 which indicates very low if any correlation between them.

There are many different approaches to handle multi-objective reinforcement learning, such as turning it into a multi-agent learning problem, learning based on linear scalarizations, policy search without learning a value function, approximate a convex coverage set to learn policies in parallel etc [17]. But here we simply adopted the learning by linear scalarization approach and used the weighted sum of the two goals as the reward. And since both factors are important and they are somewhat correlated, we decided to give both factors 50% of weight each and use the weighted sum.

And then we mapped the weighted sum of these two goals to the range of [-1, 1]. and the reward is calculated based on the mean weighted sum with the mean being 0 and others are equally spaced [11].

Discount Factor

The discount factor γ of an MDP indicates the importance of the future reward, where a higher value means a more valuable immediate and nearby future rewards. For this project the discount factor is set to 0.85, since we prefer to get a higher reward in near future instead of that in far future, meaning we intend the user to have high likelihood to return to next session and spend more effort in the current and nearby sessions compared to that in the far future sessions. The idea is that early progress tend to encourage the users to continue and be engaged in future activities, while a lower reward in the early phase might discourage the users for the future session [11] [9]. In addition, we also consider the far

future rewards because we do intend the user to spend good effort with the virtual coach and also likely to stay.

Optimal Policy

With the setting described above, we obtained the optimal policy, which can be found in table 1. One thing caught our attention is that there was repetition of optimal actions, that is the state [100] and the state [110] both had the same activity cluster as their optimal policies. And this kind of repetitions is undesired for us since we only have eight states and there is some but still limited number of activities for each cluster, and repetitions of the same activity for two or more states may have negative affect on users. Therefore, we eliminated this sort of repetitions by randomly assigning the second best action for one of these repetitive states. Only the state [010] and state [111] have preparatory activity clusters as their optimal actions and all the other states have persuasive activity clusters as their optimal actions.

And then we obtained the reinforcement learning model for our problem, which is shown in Figure 1. The virtual coach can assign an activity cluster to a user according to their state and the optimal policy. And then we obtain the reward score, which is the weighted sum of the effort spent by the user and the likelihood of their return to the next session, meanwhile the user moves to the next state.

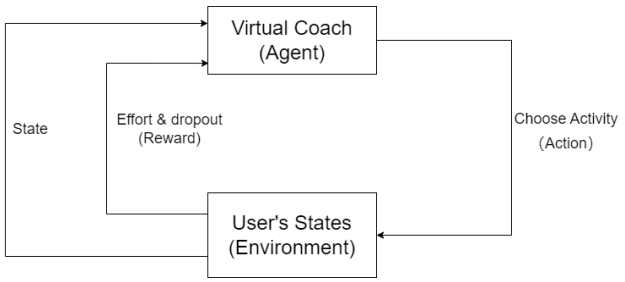

Figure 1: The reinforcement learning model with which the virtual coach chooses an activity based on the user's current state and the optimal policy then it gives the reward and also the next state of the user.

4 Experimental Setup and Results

Based on the RL model described in chapter 3, we can start investigate the analytical research questions to evaluate the effectiveness of our algorithm. In this chapter, we will go over each of these analytical questions mentioned in chapter 2 and describe their experimental setup and results.

4.1 Research Question 1

Q1: How well can states derived from the features predict behavior?

Setup. Users may behave differently in different states even though they perform the same activity. And here the behaviour, the effort the users spend on the activities and the likelihood of them returning to next session, is the two goals we intend to achieve and it is what the reward function of the RL model represents. To find out if different states indeed leads to different behaviour, we compared 1) the mean reward based on only the action with 2) the mean reward based on both the action and the state. We used leave-one-out-cross-validation, where we left out the transition samples of one person, for all 547 participants to compare these two approaches by calculating the mean L1-error for the reward prediction and its Bayesian 95% credible interval(CI) for every state.

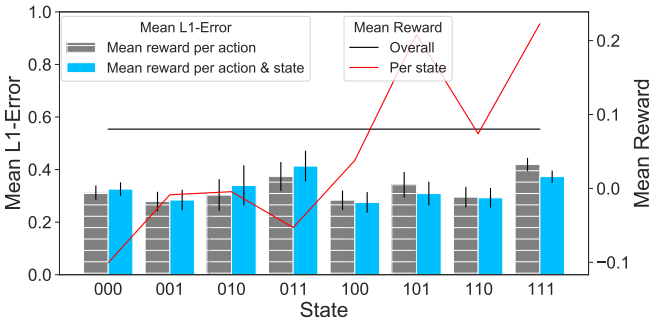

Figure 2: Left Axis: the mean prediction error per state based on 1) only the action and 2) the action and the state. Right Axis: the mean reward of each state and overall

Results. As shown in Figure 2, the average reward for all states is 0.08, and the average rewards per state varies. The state [111] and the state [101], where the users have above average beliefs that the mindset that physical activity helps to quit smoking and self-efficacy are useful, have the most highest average reward per state, with them being 0.22 and 0.21 respectively. On the contrary, the state [000] and the state [011] have the lowest average reward per state, with them being -0.1 and -0.05 respectively. As mentioned in section 3.3, state [000] means the users have below average usefulness beliefs about all three competencies, while state [011] means the users have two out of three above average beliefs of the competencies, which are self-efficacy and the awareness of smoking pattern, however, its average reward are even lower than the states who just have one out of three above average usefulness beliefs.

For most of the states adding the states to predict the behavior does not show clear differences, neither benefits nor drawbacks, from predicting the behavior based only on the actions. However, it does result in better prediction of behavior for the state [111], where the average reward for the state is the highest, than that of actions only. Therefore, we conclude that even though for most of the states we cannot draw a conclusive answer but we assume that the probability of giving better prediction based on the states and actions is higher than it offers no benefits since there is one state where adding states does offer better prediction.

4.2 Research Question 2

Q2: How well can the states derived from the state features predict states after choosing activities?

Setup. By choosing the optimal activities, we intend to get the users to a state where they spend more effort on the activities and are more willing to come back for the next session. Therefore we need to be able to predict the next states of the users after they perform the proposed activities. To evaluate how well our model can predict the next states, we used the leave-one-out cross validation to compare three ways of giving such predictions: 1) assigning equal probability for all states, 2) predicting the users' next states as the current states they are in, and 3) calculating the probability based on the transition function of our model. And the comparison is based on the mean likelihood of each next state and their 95% CIs.

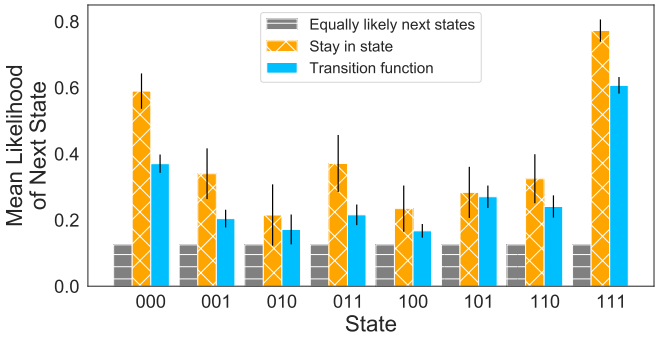

Figure 3: The mean likelihood of a transition to a next state by 1) assigning equal probability to all next states, 2) assuming stay in the current state, and 3) using probability based on estimated transition function of RL model

Results. As shown in Figure 3, both the prediction of the next states of the users by assuming they stay in the current state and the prediction of the next state of the users by using the probability based on the estimated transition function of our model result in much higher mean likelihood of next states than that of assigning equal probability for every state. This means that the probability of the next states is not a uniform distribution. For states [000], [001], [011] and [111], the mean likelihood of the next states from the three approaches do not have overlaps with their CIs and among them states [000] [011] and [111] have the highest mean likelihood of next states when predicting the next states to be the current ones. Considering the mean reward per state, it means that the users tend to stay in their current states when they are in the states that they are either spending more effort and willing to stay for next session or the very opposite of this. In a word, considering people's current states indeed improve the prediction of the next state. But people who either spend very little or very much effort in the activity and either have very strong or very little motivation to return to the next session are most likely to stay in their current states. And this could mean that our RL model might not be able to move some of the least motivated users to a better state.

4.3 Research Question 3

Q3: What is the effect of performing multiple optimal chosen activities on the users' states?

Setup. After checking how well the model can predict the next states after one action, we can look into the effect of the optimal actions on the states in the future, since our goal is to move the users to the states that have higher rewards. In order to see the effect, we first used the data to gain the optimal policy $\pi^\star$, which was calculated by value iteration on the aforementioned data. And then we simulated users with even distribution of all states to follow $\pi^\star$ for some certain number of actions and checked the distribution of the users in all states after that.

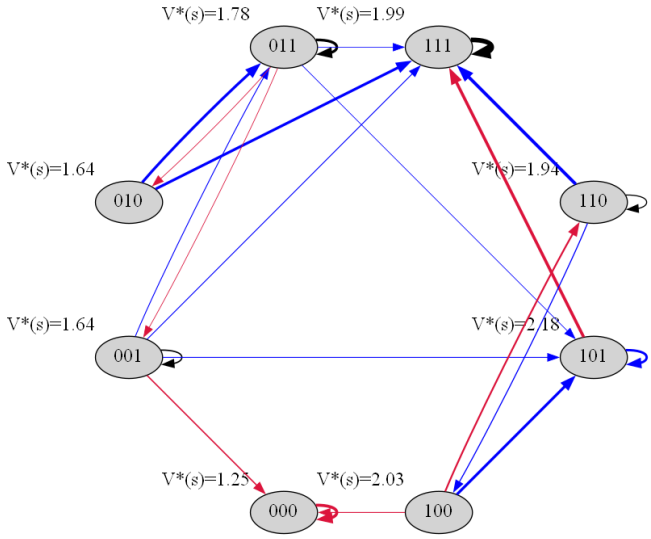

Figure 4: Transition probabilities under $\pi^\star$. Only transition with a probability of at least $\frac{1}{|S|}$ are shown. Red arrows means transitions to next states with a lower value. Black arrows means transitions to states with the same value. Blue arrows means transitions to states with higher value. The thicker the line, the higher the probability of the transition.

Results. Figure 4 shows the transition functions under the optimal policy $\pi^\star$ when they are at least $\frac{1}{|S|}$ and also the cumulative discounted reward $V^\star(s)$ for each state. From the network we can see that, users tend to move to the states with higher $V^\star(s)$ such as states [011], [111], [110] and [101]. Moreover, once they have reached those higher value states, they somewhat tend to stay in those states, such as 44% probability of stay at state [101] and 86% for state [111]. But not always, for example, users in state [100], where the value is the second highest, 2.03, and the users have above average belief of only one competency, the mindset that physical activity helps to quit smoking, are less likely to stay there or move to a better state. In addition, the state with highest value is not [111], with 1.99 but [101], with 2.18. However, the red arrows in the network also indicates that there are still chances of the users moving from a higher value state to a lower one, and among them a potential drawback of our model is that when the users are in state [000], where the $V^\star(s)$ is the lowest, they tend to stay there with a probability of 55% and the state [100] can also lead to this state.

Figure 5 shows the distribution of people in each state af-

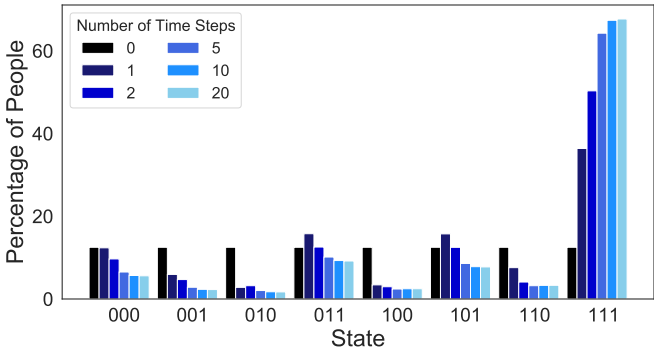

Figure 5: Percentage of people in each state after following $\pi^\star$ for some certain number of steps.

ter following the optimal policy $\pi^\star$ for, not just one step, but some certain number of steps up until twenty steps to illustrate the long-term effect of the optimal policies on the users' states. From the graph we can see that, in general, people moved from the other states to state [111], where average reward per state is the highest and the accumulative discounted reward is relatively high. And this is reasonable since the activities are supposed to make the users more aware of the usefulness of the three competencies. However, there are still some users in the states where the cumulative discounted reward is low. Especially state [000] still has 5.6% of all users. In conclusion, The model can move most of the users to the state where they spend the most effort and are most likely to stay for the next session while it fails to move some users to better states where they spend more effort and are more likely to come back to the next session.

4.4 Research Question 4

Q4: How effective is the optimal activities compared to non-optimal ones in their effect on behavior?

Setup. Now we evaluated the effect of the optimal policy on the immediate states and the future states of the users, we would like to further evaluate whether the optimal policy is indeed better than non-optimal ones, judging by the behaviour of the users. In order to test it, we compare three policies 1) the optimal policy 2) the average policy, and 3) the worst policy, by their mean reward on a series of certain timesteps up until 100. Here the average policy is the policy that makes sure each action was taken an equal amount of times at all timesteps and the worst policy is the least optimal policy, the action that has the least reward, that we gained by training the data. We initialize the distribution of the users in different states for this experiment in two ways. One way is to initialize the distribution according to the distribution of all the people in the first session from the data. This is is to see if the model can indeed achieve our goals in general, which is to get the users to spend more time and be more willing to participant in these sessions. On top of that, the other way is to initialize only based on the state distribution of the people who have lower than 25% of the reward among all the people in the first session from the data. And this is to see how well the optimal policy can perform for the people who need the help the most.

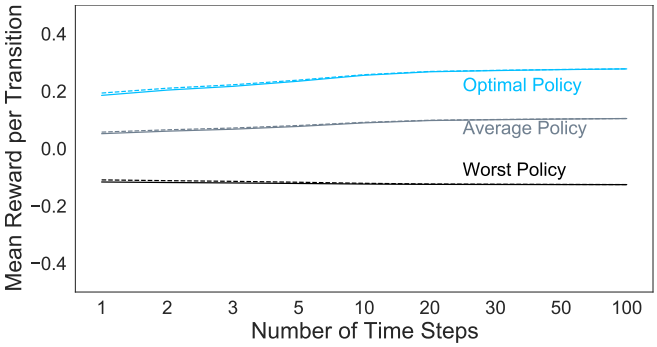

Figure 6: Enter Caption

Results. Figure 6 shows the mean rewards per transition for these three policies at different number of time steps. From the plot we can see that the optimal policy results the best mean reward per transition for both the general users and the users who have the least reward at the beginning. For the general population, the optimal policy raised the mean reward from 0.2 to 0.28, at 20 steps, and to 0.29 at 100 steps. For the least well-performed population. the optimal policy also managed to increase the mean reward from 0.21 to 0.29 by 0.08. The average did manage to increase the mean reward, but only by 0.05. Therefore, out model does perform better than the average policy and the worst policy, regarding to optimize activities so that the users can make more effort and be more intrigued to go back to the sessions.

4.5 Research Question 5

Q5: How do different optimal policies based on different reward functions compare with each other regarding the effect of multiple optimal actions on behavior?

Setup. Finally, there is one last factor we wanted to evaluate, our reward function. As mentioned before, we used weighted sum of the two objectives, the user's willingness to return for next session and their effort spent on the proposed activities, we wanted to see how weights affect our model. We compared the model of three different weights, 1) the sum of 50% effort and 50% return rate, which is our original model, 2) the sum of 25% effort and 75% return rate which gives more weight to the return rate, and the opposite 3) the sum of 75% effort and 25% return rate, which gives more weight to the effort spent. First of all, we wanted to see if different weights lead to different optimal policies. And then we further investigated the effectiveness of the different optimal policies to move the users to better states. Here it is worth mentioning that we do not select the features based on different weights again, instead we continued to use the initial selected features to form the same state space. By doing so we hope to evaluate the robustness of the reward function without changing too much factors at once.

Results. As shown in Table 1, the optimal policies for different approaches are not always same for the states. For ex-

Weight	[000]	[001]	[010]	[011]	[100]	[111]	[110]	[111]
0.25	9	12	2	5	13	4	13	7
0.5	9	12	1	5	13	4	13	7
0.75	5	12	3	5	13	4	7	6

Table 1: Optimal policies obtained from different weighted sum of the effort and the return likelihood by effort being 1) 25%, 2) 50% and 3) 75% of the sum.

ample, in state [010] the optimal policies for the three approaches are all different from each other. But the optimal policies for the rest of the states are identical for at least two approaches. States [001], [011], [100] and [111] have the same optimal policies for all three approaches. In addition, only state [010] and state [111] have preparatory as their optimal actions and all the others have persuasive activity clusters as their optimal actions.

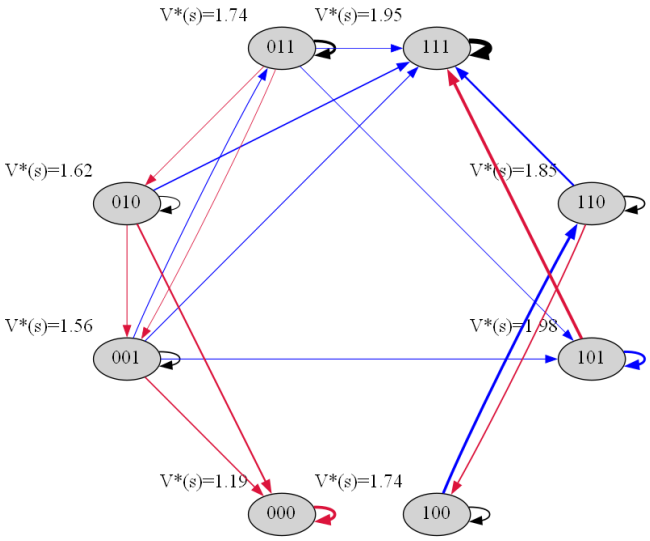

Figure 7: Transition probabilities under the optimal policy obtained by effort being 25% of the weighted sum. Only transition with a probability of at least $\frac{1}{|S|}$ are shown. Red arrows means transitions to next states with a lower value. Black arrows means transitions to states with the same value. Blue arrows means transitions to states with higher value. The thicker the line, the higher the probability of the transition.

Figure 7 and Figure 8 show the transition functions under the optimal policies, which were obtained with the effort being 25% and 75% of the weighted sum for reward respectively, when they are at least $\frac{1}{|S|}$ and also the cumulative discounted reward $V^{\star}(s)$ for each state. Combining with the plot 3 of our original weighted sum where effort spent and return likelihood being equally important, we can see that all three approaches have somewhat similar transition networks. In detail, all the three optimal policies tend to move the users to better states, and also keep them in those states. And they all have some red arrows that indicate they may not be able to move the population in some bad states to better states.

However, there are some differences worth mentioning. First, their values for different states are not always the same. In fact, the average value of the 75% effort approach is the

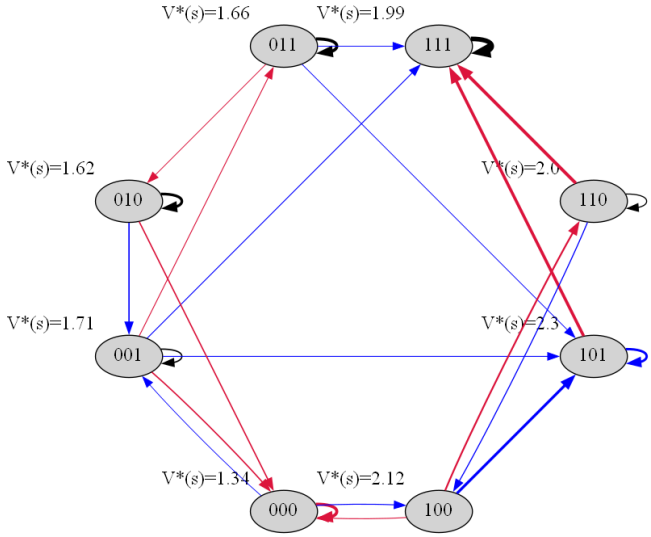

Figure 8: Transition probabilities under the optimal policy obtained by effort being 75% of the weighted sum. Only transition with a probability of at least $\frac{1}{|S|}$ are shown. Red arrows means transitions to next states with a lower value. Black arrows means transitions to states with the same value. Blue arrows means transitions to states with higher value. The thicker the line, the higher the probability of the transition.

highest, 1.84, with that of the 50% effort approach and that of the 25% effort approach being 1.80 and 1.70 respectively. The 75% effort approach also has the highest value, 2.3 and lower lowest value, 1.34. Besides, the optimal policy with effort being 75% of the weighted sum has better chance to get the users out of the worst state [000]. In general, the approach with effort being 75% may be the best of these three regarding their transition functions and accumulative reward.

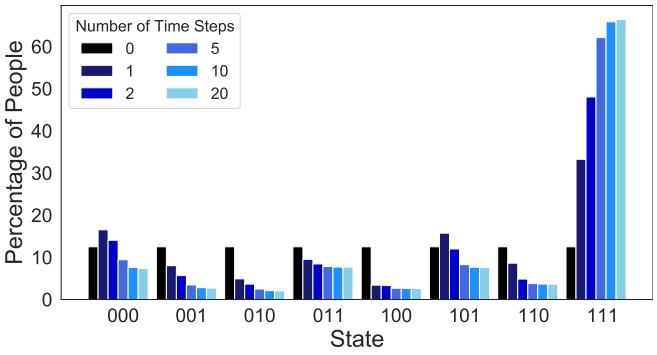

Figure 9: Percentage of people in each state after following the optimal policy obtained with effort being 25% of the weighted sum for some certain number of steps

Furthermore, Figure 9 and Figure 10 show the distribution of the population for multiple different timesteps in each state after following the optimal policies with the effort being 25% and 75% of the weighted sum respectively. Just like the model with the original reward function, both approaches were able to move the population to the state [111], where the

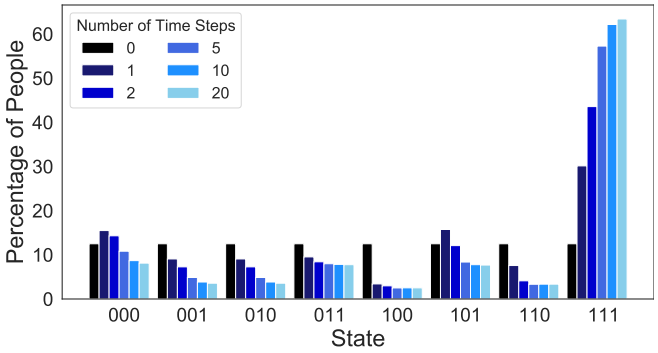

Figure 10: Percentage of people in each state after following the optimal policy obtained with effort being 75% of the weighted sum for some certain number of steps

average reward per state is the highest and the accumulative discounted reward is relatively high. In addition, the original policy was better at decreasing the distribution of the population in the worst states. For example, it lowered the population in state [000] to 5.6% while the other two approaches only managed to decrease it to around 8%. Therefore, our original weight assignment perform better than the other two, which means the both the effort spent and the return likelihood are important.

5 Responsible Research

This research was done with the principles of the responsible research, including the reproducibility of the methods used and also the ethical aspects of the research.

Reproducibility. The paper itself provides clear explanation about the methods used including the RL algorithm illustrated in chapter 3 and also the setups for the evaluation experiments in chapter 4. The code for the paper is also provided and organized in a clear way with a guideline, including the organization of the code, the purpose of each file, and the references used. Besides, files and methods were well-documented to help readers to understand what to expect from these files and methods. In conclusion, by following the guidance and documentation of the code readers with computer science background should be able to understand the code and reproduce the results if the data is available.

Ethical Aspects of the Research. During the process, we always kept in mind the principles of carrying out ethical research. For starters, some may wonder if it is truly ethical to develop something that persuade people to do certain tasks, even it is for quitting smoking. To our defense, the study for the data collection with the virtual coach was approved by the Human Research Ethics Committee of Delft University of Technology [13]. On top of that, the pre-screening process had excluded people who did not give consent, who was under the age of 18 and who was not contemplating or preparing to quit smoking at that time and in this way the eligible participants are the ones who can give and gave informed consent and whose goal was inline with the study. Therefore there was no forceful or non-consented activities during the study. In addition, there was no untruthful data processing during the study to skew the data to match our expectations and we always tried to interpret the data truthfully. Therefore, the results should be true and honest.

6 Discussion and Conclusion

Contributions. The paper evaluated the effectiveness of using a reinforcement learning model to suggest activities for preparing to quit smoking. From our experiments we found that adding states is more likely to result in better predictions of the users' behavior, and it is indeed better for one state [111], where the average reward per state is the highest (Q1). The average reward is highest at state [111] and [101], but the state [011] has the second lowest average reward than the states where there is only one above average usefulness beliefs (Q1). And the power of using states to predict the next states depends on the states, more specifically people tend to stay in better states and also in the worst states (Q2). In addition, with the activities chosen according to the RL model, people tend to move to better states and also stay in better states where they spend more effort on the activities and are more willing to return, despite that there are still some people who stay in the less good states (Q3). And this finding is also in agreement with the fact that the optimal policy does result in better reward in the long run for all the users and for the users who need the most help (Q4). At last, comparing the effect of changing weight distribution of the two objective while calculating the reward function, we can conclude that the initial distribution of the weight is a reasonable choice (Q5). Different policies derived from different reward functions are somewhat similar but still have differences. Besides, we can see that most of the optimal actions for all three cases are persuasive activities.

Limitations and Future works. Despite the supportive evidence that this paper provides for using RL in this context, there are some limitations of our results. First of all, the reward function we use is based on the weighted sum of the two goals, however, this might not be the best and the most accurate way to describe the relation between these two goals. So one direction for future works could be to investigate the relation between these two goals and to have a RL model whose reward function is formulated in a different way and compare with this one. Secondly, in our experiments we assumed that the transition function and the reward function do not change over time steps. However, this may not be the case for real life. For example, repetitions of the same activity clusters may help to build the users identity [14] and thus motivates them to spend more effort on the sessions. And it could also be the opposite case. Lastly, there is some potential inconsistencies in our results. One is that the state with the highest accumulative discounted reward is not [111] but [101] for all three cases, as showing in graph 4, 7, and 8. The reason for this could be that believing the awareness of smoking patterns useful has negative impact on the users' partici-

pation of the activities, but it could also be that the data we used. The other is that there are still possibilities that users can go from a better state where they have above certain average usefulness beliefs to a state where they have below average usefulness belief about that certain competency. For example, in graph 4, there are transitions from state [100] to state[000] where the users' belief that the mindset that physical activity helps to quit smoking is useful decreased. These consistencies could be that we tested our model with leave-one-out cross-validation simulations, instead of testing on human subjects due to limited budget, and it may not be able to accurately show the actual affect of our model. Therefore, some future tests can be carried out to evaluate the model.

Conclusion. We wanted to see if a reinforcement learning model can be used to choose activities for preparing to quit smoking, to be more specific, we wanted to evaluate how effective such a model can be to optimize two goals, the likelihood of the users return to the next session and also the effort that the users spend on these session, so that they can make the most use of these sessions. And based on our implementation and evaluation, such a RL model is indeed effective when it comes to improving these two goals and moving the users to a relatively better, where they tend to make a good effort and are willing to stay for future session.

7 Acknowledgement

This work is part of the multidisciplinary research project Perfect Fit, which is supported by several funders organized by the Netherlands Organization for Scientific Research (NWO), program Commit2Data - Big Data & Health (project number 628.011.211). Besides NWO, the funders include the Netherlands Organisation for Health Research and Development (ZonMw), Hartstichting, the Ministry of Health, Welfare, and Sport (VWS), Health Holland, and the Netherlands eScience Center.

References

[1] Babb S; Malarcher A; Schauer G; Asman K; Jamal A. Quitting smoking among adults — united states, 2000–2015. *MMWR Morb Mortal Wkly Rep*, 2015. http://dx.doi.org/10.15585/mmwr.mm6552a1.

[2] Steven A and Schroeder M.D. We can do better — improving the health of the american people. *New England Journal of Medicine*, 357(12):1221–1228, 2007. https://doi.org/10.1056/NEJMsa073350.

[3] Nele Albers. Virtual Coach Mel for Proposing Useful Preparatory Activities for Quitting Smoking.

[4] Emily Banks, Grace Josh, Marianne F. Weber, Bette Liu, Robert Grenfell, Sam Egger, Ellie Paige, Alan D. Lopez, Freddy Sitas, and Valerie Beral. Tobacco smoking and all-cause mortality in a large australian cohort study: findings from a mature epidemic with current low smoking prevalence. *BMC Med*, 2015. doi: 10.1186/s12916-015-0281-z.

[5] David Chapman and Leslie Pack Kaelbling. Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. *In Proceedings ofthe 12th International Joint Conference on Artificial Intelligence (Sydney, New South Wales, Australia), John Mylopoulos and Raymond Reiter (Eds.)*, 1991.

[6] Cooper J, Borland R, Yong HH, McNeill A, Murray RL, O'Connor RJ, and Cummings KM. To what extent do smokers make spontaneous quit attempts and what are the implications for smoking cessation maintenance? findings from the international tobacco control four country survey. *Nicotine Tob Res.*, 2010.

[7] Siddiqi K., S. Husain, Vidyasagaran, and A. et al. Global burden of disease due to smokeless tobacco consumption in adults: An updated analysis of data from 127 countries. *BMC Medicine*, 2020. https://doi.org/10.1186/s12916-020-01677-9.

[8] López-Torrecillas F.; Ramírez-Uclés I.; Rueda M.d.M.; Cobo-Rodríguez B.; Castro-Martín L.; Urrea-Castaño S.A.; Muñoz-López L. Use of the therapy app prescinde for increasing adherence to smoking cessation treatment. *Healthcare*, 2023. https://doi.org/10.3390/healthcare11243121.

[9] Amabile Teresa M. and Steve J. Kramer. The progress principle: Using small wins to ignite joy, engagement, and creativity at work. *Harvard Business Review Press.*, 2011.

[10] McGinnis J. Michael and William H. Foege. Actual Causes of Death in the United States. *JAMA*, 270(18):2207–2212, 11 1993.

[11] Albers Nele, Neerincx Mark A., and Brinkman Willem-Paul. Persuading to prepare for quitting smoking with a virtual coach: Using states and user characteristics to predict behavior. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 717–726, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[12] Albers Nele and Brinkman Willem-Paul. Perfect fit - beliefs about and competencies built by preparatory activities for quitting smoking and becoming more physically active. 2022.

[13] Albers Nele and Brinkman Willem-Paul. Perfect fit - learning to propose useful preparatory activities for quitting smoking and becoming more physically active. 2023.

[14] Kristell M. Penfornis, Winifred A. Gebhardt, Ralph C.A. Rippe, Colette Van Laar, Bas van den Putte, and Eline Meijer. My future-self has (not) quit smoking: An experimental study into the effect of a future-self intervention on smoking-related self-identity constructs. *Social Science Medicine*, 320:115667, 2023.

[15] Doll R, Peto R, Boreham J, and Sutherland I. Mortality in relation to smoking: 50 years' observations on male british doctors. *BMJ*, 2004. https//10.1136/bmj.38142.554479.AE.

[16] Hannah Ritchie and Max Roser. Smoking. *Our World in Data*, 2023. https://ourworldindata.org/smoking.

[17] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, A Whiteson Nl, and Richard Dazeley. A survey of multi-objective sequential decision-making, 2013.

[18] Kelders S, Kok R, Ossebaard H, and Van Gemert-Pijnen J. Persuasive system design does matter: A systematic review of adherence to web-based interventions. *J Med Internet Res*, 2012. DOI: 10.2196/jmir.2104.