Using Reinforcement Learning to Determine When to Provide Human Support in Quitting Smoking with a Virtual Coach

Shirley Li

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

[1]**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
June 23, 2024

Name of the student: Shirley Li
Final project course: CSE3000 Research Project
Thesis committee: Nele Albers, Willem-Paul Brinkman, Zhengjun Yue

Abstract

Smoking remains one of the largest health concerns worldwide, which is why eHealth applications with virtual coaches have been developed to assist smokers with quitting. Providing additional feedback from human coaches during such smoking cessation programs can further improve the effectiveness of the intervention. However, due to budgetary constraints and the limited availability of human coaches, it is important to make informed decisions about when someone gets human support to optimize the effectiveness. This research investigates the use of reinforcement learning (RL) to determine when to provide human feedback in quitting smoking with a virtual coach. Using data from a longitudinal study, we implemented an RL model that decides when to involve a human coach based on users' appreciation for human support and their self-efficacy, optimizing the effort that people spend on preparatory activities and their likelihood of returning to the program. Results show that the model is effective in allocating human support, increasing users' effort and return likelihood while considering the cost of human coaches. These findings support using RL to help with determining when to provide human support in smoking cessation programs.

1 Introduction

In the Netherlands, 19% of people aged 18 years or older indicated to smoke and over a third of those people smoke every day (13.5% of the whole population)[1]. More than half of the people that smoke regularly, die from smoking-related diseases, such as cancer or chronic lung disease. This comes down to roughly 19.000 people in the Netherlands per year [1] and 8 million people worldwide [2]. Smoking is highly addictive and it may take 30 or more attempts to be successful in quitting smoking [3]. Next to that, a study has shown that smoking cessation with assistance leads to higher success rates than without [4]. A popular way of assisting with behavioral change therapies, such as quitting smoking, is using an eHealth application. Such applications utilize digital technologies to support users in improving their health status.

In recent years, eHealth applications have become more and more available to help people change unhealthy behavior, such as smoking or obesity. Studies have shown that these applications can have positive effects on users and help them with achieving their goals (e.g. [5], [6]). Many eHealth applications use a virtual coach to ask patients questions about their health status and then provide feedback or suggestions to improve. However, these applications rely solely on a virtual coach and do not also consider a human coach that can intervene when needed. Chikersal et al. [7] showed that human support can have positive effects on the outcome of health interventions and Lee et al. [8] showed that users who interacted with a system that also offered human support showed more engagement and trust for the system compared to users who did not get human support. Albers et al. [9] also found that people appreciate the situational awareness, empathy, and accountability that a human coach brings to a smoking cessation program. Thus, integrating human support with eHealth applications could be beneficial during behavior change interventions. However, due to a limited amount of money or time for human coaches, it might not be feasible to provide human support for everyone. Therefore, it is crucial to be able to make informed decisions on when to offer human support to maximize the effectiveness with the limited amount of resources. While there is research on how to incorporate human support in eHealth interventions (e.g. [10]), there is limited research on how to determine when to involve a human coach.

There are multiple possible strategies to determine when to give human feedback to a user. For example, it could be done randomly or it could be provided to new users only. However, each person has different wants and needs, and the way they feel can vary daily. Therefore, a personalized approach is preferred, which is in line with behavior change theories [11]. Factors such as how a person feels and their desires can be considered as the state that they are in. Based on this state, we would want to determine whether or not to provide human feedback to make optimal tailored decisions. For example, Albers et al. [12] showed the importance of considering users' states when choosing a persuasion type to quit smoking. On the other hand, receiving human support can also influence the future state of a person. For example, their motivation might increase if they receive human support [10]. Therefore, we also need to consider the consequences of providing human support in determining when to do it.

One framework that takes people's current state and future states into account when making decisions, is Reinforcement Learning (RL). Several studies have already used RL to improve the effectiveness of eHealth applications (e.g. [13], [14]) and Weimann et al. [15] confirmed the potential and effectiveness of using RL for behavior change interventions by a systematic literature review. Piette et al. [16], [17] and Forman et al. [18] have also shown that using Reinforcement Learning to determine whether to provide a message from a human coach or an automated message can be effective and reduce the work for therapists. However, both of these studies use reinforcement learning to select one of the two options, human feedback or AI-generated feedback, and do not use human support as an additional service. Furthermore, Forman et al. [18] did not consider the state that a user is in, but only their behavior (e.g. meeting physical activity goals and weight loss) for the RL model. Therefore, this research paper will explore the use of an RL model that considers both the users' state and behavior, to determine when to provide human feedback in addition to quitting smoking with a virtual coach. The following research question will be answered in this paper:

How effective is a reinforcement learning model in determining when to provide human feedback that optimizes the effort people spend on their activities and the chance that they stay in the intervention?

This research uses data from a study by Albers et al. [19]

where 852 smokers/vapers interacted with a chatbot Kai in up to five sessions to train a reinforcement model in determining when to provide human feedback. In each session, Kai first asked questions to determine people's state and then suggested a preparatory activity for quitting smoking. At the next session, the chatbot asked about the effort they spent on the activity and the likelihood of returning to the intervention if it was an unpaid smoking cessation program. In between sessions, participants had a 20% chance of receiving feedback from a human coach. Using all this collected data, we implemented a reinforcement model that optimizes the effort that people spend on the activity and the likelihood of returning to the intervention while considering the cost of a human coach. We then measured the effectiveness of predicting users' behavior (effort and likelihood of staying in the intervention) and future states. We also used simulations to assess the long-term effects of the RL model and to assess different reward functions for the model.

This research contributes to the scientific understanding of the effectiveness of using reinforcement learning in eHealth applications for behavior change interventions. It demonstrates the potential of RL for improving users' behavior and supports the use of RL to help with determining when to provide human feedback in smoking cessation programs. Using this, we can work towards more effective interventions where the availability of digital technologies is combined with knowledge, empathy, and accountability from human coaches.

2 Methodology

To answer the research question, we developed a reinforcement learning model that determines when to provide human support in preparing to quit smoking. The model aims to optimize the effort that users spend on the preparatory activity that the virtual coach suggests and the likelihood of returning to the intervention. To develop the model, we used data collected from a study by Albers et al. [19] and closely followed the approach from Albers et al. [13] for a similar study. The following section describes the virtual coach, the human support, the reinforcement model, and how the data is collected.

2.1 Virtual Coach

During the study, users interacted with a text-based virtual coach Kai [20] that was implemented to help users prepare for quitting smoking during five sessions. In each session, Kai first asked users about their self-efficacy, the importance of preparing for quitting smoking, how they viewed getting human feedback, and their energy level. The virtual coach then suggested a preparatory activity for quitting smoking (e.g. tracking one's smoking behavior). In the next session, Kai inquired about users' experience with the activity, how much effort they spent on it, and how likely it is that they would have returned to the sessions if it was an unpaid smoking cessation program.

2.2 Human Support

Between the sessions with the virtual coach, participants had a 20% chance of receiving personal feedback on their progress from a human coach. Two human coaches with a background in psychology wrote these messages and the participants received them through Prolific. To help the coaches write the feedback messages, each participant wrote a short introduction text at the start of the study. In this introduction, users could mention matters like their motivation to quit smoking, previous experiences with quitting, and areas where they need help. In addition to the introduction text, the coaches had access to users' baseline smoking/vaping frequency, baseline weekly exercise amount, and the information from the virtual coach from the last session. The messages that the coaches wrote were based on a paper by Ghantasala et al. [21] and consisted of feedback on their progress, suggestions on how to improve, and the reasoning behind these suggestions.

2.3 Data collection

Study. The data is gathered from a longitudinal study conducted by Albers et al. [19] where participants interacted with the virtual coach Kai in up to five sessions. The study was approved by the Human Research Ethics Committee of the Delft University of Technology (Letter of Approval number: 3683). The participants were recruited from the online crowdsourcing platform Prolific. To be eligible for the study, people had to smoke/vape daily, be over the age of 18, be fluent in English, be contemplating or preparing to quit smoking, not be in another smoking/vaping cessation program, and provide informed consent.

Data. There were 852 eligible people who started with the study and we collected 2323 $\langle s, a, r, s' \rangle$-samples from 678 people, where s is the state, a the action, r the reward, and s' the next state. These samples only contained data from people who went to at least 2 sessions and answered all the relevant questions. The characteristics, such as age and gender, of these 678 people can be found in Table 2 in the Appendix.

2.4 Reinforcement learning model

To answer our research question, we implemented a reinforcement learning model that learns a policy to determine when to provide human support such that the effort people spend on their preparatory activities and their likelihood of staying in the intervention are maximized. This can be formally described as a Markov Decision Process (MDP) $\langle S, A, R, T, \gamma \rangle$, where:

- S is the state space,
- A is the action space,
- $R(s, a, s')$ is the reward for transitioning from state s to s' by taking action a,
- $T(s, a, s')$ is the probability of transitioning from state s to s' by taking action a,
- and γ is the discount factor.

In an MDP, the goal of the agent is to learn an optimal policy π^* that maximizes the expected cumulative discounted reward over time $\mathbb{E}[\sum_t^\infty \gamma^t r_t]$, where t denotes the time step, γ^t the discount factor at time step t and r_t the reward at time step t. Such a policy describes the set of optimal actions to take in each state. To compute the optimal policy, we used policy iteration where we evaluated the policies with their Q-value function.

States

The set of states S consists of a subset of the personal features that the virtual coach Kai asked about. During each session, users were asked about their self-efficacy (e.g. how confident they are in preparing to quit smoking) based on McAuley's Exercise Eelf-Efficacy Scale [22], how important they think it is to prepare to quit smoking based on a question by Rajani et al. [23], how much energy they have and how they would view getting human feedback after the session. All of these were on a scale from 0 to 10, except for the last question which was on a scale from -10 to 10.

To ensure we had enough state-action pairs in our data to train the model, we first needed to shrink the state space by converting the feature values in their original scales to a binary or tertiary representation. We also selected a subset of the state features to further reduce the number of states. To select the features that we wanted to use for our model and how many possible values they could take, we adapted the G-algorithm from Chapman and Kaelbling [24] in a way that was inspired by the methods used by Albers et al. [13] and Dierikx [14]. This algorithm iteratively selects the feature for which the Q-values are most different for each of the possible values that the feature can take. We adapted it to use an ANOVA test instead of a t-test to test for the differences in Q-values to accommodate for possible tertiary features. The selected features and their number of possible values were 1) appreciation of human support with three possible values, and 2) self-efficacy with three possible values. To map the original feature values to their corresponding representation, we looked at the percentiles of each feature. For example, for a binary representation, we calculated the 50th percentile and mapped everything below or on the 50th percentile to 0 and everything above to 1. This resulted in a state space of $3 * 3 = 9$. States are denoted as strings such as 21, here the first feature is a 2, denoting a relatively high want for human support, and the second feature is a 1, which indicates a relatively medium self-efficacy.

Actions

The actions that can be taken by the agent are either providing human support or not. This is denoted by a 1 and 0 respectively.

Reward

To compute the reward, we used the data from sessions 2-5 where users were asked about the effort they spent on the proposed preparatory activity on a scale from 0 to 10, adapted from Hutchinson and Tenenbaum [25], and the likelihood of returning to the intervention if it was an unpaid smoking cessation program on a scale from -5 ("definitely would have quit the program") to 5 ("definitely would have returned to this session").

To calculate the reward, we first transformed the return response scale from -5 to 5 into a 0 to 10 scale by adding 5 to each value. Then, for both effort and return responses, we computed the individual reward $r \in [-1, 1]$, by mapping the mean response to 0 and linearly scaling the responses below and above the mean to a range of -1 to 1. This follows the approach by Albers et al. [13] and is done by the following function:

$$r = \begin{cases} -1 + \frac{x}{\bar{x}} & \text{if } x < \bar{x} \\ 1 - \frac{10-x}{10-\bar{x}} & \text{if } x > \bar{x} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here $\bar{x}$ is the mean of the responses and x is the to-be-mapped value.

Now we have two separate reward values that represent the two objectives. To combine these objectives, there are multiple possible strategies, such as using a linear scalarization function to turn it into a single-objective problem, or using more complex multi-policy approaches [26]. A simple linear scalarization function seems to be the most used approach for RL in behavior change interventions [15], which is why we opted for the weighted sum method to combine our objectives. This also follows the approach used by Piette et al. [16], who composed their reward of the two objectives in equal proportions. We also used equal weights of 0.5 for our effort and return rewards, such that they are equally important. Next to weighing both the effort and return responses, we also wanted to account for the cost of involving a human coach. Therefore, we subtracted a cost factor of 0.21 if the action of providing human support was chosen. This number was chosen, because this results in providing human support about 4% of the time, which is closest to the 20% chance in the study. The final reward function was thus:

$$R(s, a, s') = 0.5r_e(s, a, s') + 0.5r_r(s, a, s') - 0.21a$$

Here s is the current state, a is the chosen action, s' is the next state, and $r_e(s, a, s')$ and $r_r(s, a, s')$ are the separate rewards for the effort and the return ratings provided in state s', respectively.

Discount factor

The discount factor γ was set to 0.85, following the approach of Albers et al. [13] to favor rewards in the near future versus the distant future. This ensures that we prioritize actions that have a positive effect early on in the intervention, which might prevent people from dropping out. However, we also still want to account for long-term success, so we set the discount factor γ to 0.85.

3 Evaluation Setup and Results

To answer the research question, we have five subquestions that help analyze the effectiveness of our reinforcement learning model. This section will answer the subquestions by describing the evaluation setup followed by an overview of the results.

Q1: How well can the states derived from the provided user data predict users' behavior after receiving human support?

Setup. The state that a user is in might help with predicting their behavior after receiving human support. The behavior is in our case the effort they spend on the preparatory activity and the likelihood of returning to the intervention, which is represented in the reward function of our reinforcement model. To measure the effectiveness of predicting the reward considering the states, we compared two different approaches of predicting the reward: 1) the mean reward per action and 2) the mean reward per action and state. To compare these approaches, we calculated the mean L_1-error and its Bayesian 95% credible interval (CI) per state using leave-one-out cross-validation for the 678 participants with at least one transition sample. With leave-one-out cross-validation, we leave the data from one person out from the training data and compare the prediction for the left-out person with the actual value. If two 95% CIs are not overlapping, we have credible information that the two values are different.

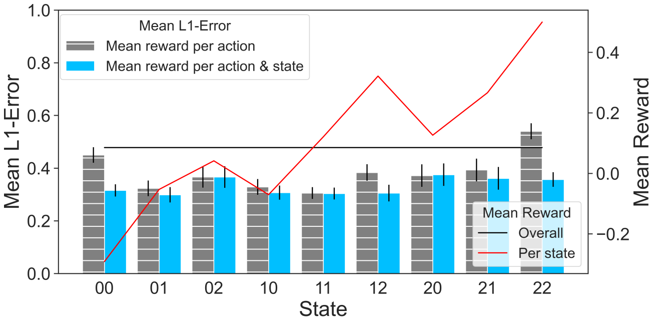


Figure 1: Left axis: Mean L_1-error with 95% CIs for reward prediction based on 1) the mean reward per action, and 2) the mean reward per action and state. Right axis: Mean reward overall and per state.

Results. Figure 1 shows that considering both action and state tends to result in lower L_1-errors than only considering the action for predicting the reward. This is especially the case for the states where the reward is much higher or lower than the overall mean reward (states 00, 12, and 22). In these states, the 95% CIs do not overlap, meaning it provides credible information that considering the states for reward prediction performs better. For the other states, where the mean reward is closer to the overall mean reward, the 95% CIs do overlap, so we cannot say that one approach performs better.

Q2: How well can the states derived from the provided user data predict future states after receiving human support?

Setup. Ideally, we want the timing of providing human support to have a positive impact on the state that a user is in, i.e. we want them to get to a state where they spend more effort on the activities and are more likely to continue with the program. Therefore, we need to be able to predict the future states of a user after receiving human feedback. With the data from the transition samples, we used leave-one-out cross-validation to compare three approaches of predicting the next states: 1) equal probability of transitioning to each state, 2) staying in the current state, and 3) using the transition function that was obtained from the training data. We calculated the mean likelihood of transitioning to the correct next state using each approach and its 95% CI per state.

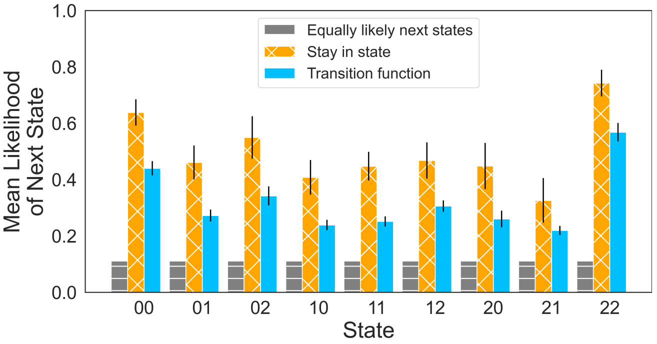


Figure 2: Mean likelihood of transitioning to the next state with their 95% CIs based on three different approaches.

Results. As one can see in Figure 2, staying in the current state or transitioning according to the estimated transition function leads to a higher mean likelihood of the next state than assigning an equal probability to all states. It can also be seen that for all states, predicting that people stay in the current state leads to the highest mean likelihood, with non-overlapping 95% CIs. This means that people tend to stay in their current state if we were to provide human feedback at random like in the study. This is not ideal, since we would want to move people from states with lower rewards to states with higher rewards. However, considering the current state works well for predicting future states and can thus help with choosing the right action such that people are moved to states where the expected reward is higher.

Q3: What is the effect of (multiple) optimal actions on users' states?

Setup. Following the optimal policy (see Table 1), we want people to ultimately get to states where they spend a lot of effort on their activities and where they are likely to stay in the intervention. To see if the optimal policy π^* that the reinforcement learning model learned, has this desired effect, we calculated the percentage of people in each state after following π^* for a various number of time steps. We started with an equal distribution of people across the states.

Table 1: Optimal policy π^* for each state

State	00	01	02	10	11	12	20	21	22
π^*	0	0	0	0	0	0	1	0	0

Results. Table 1 shows that only people in state 20 get human support according to the optimal policy. In this state, people have a high appreciation for human support and a low self-efficacy. The transition diagram in Figure 3 displays that following the optimal policy π^*, people in this state are likely to be moved to states 21 and 22 where their self-efficacy and the expected cumulative reward over time V^* is higher. However, they also have a probability of going to states 10 and 00, where V^* is lower. Nevertheless, the probability of going to a worse state is lower than going to a better state, so providing human support in this state seems to be effective. The

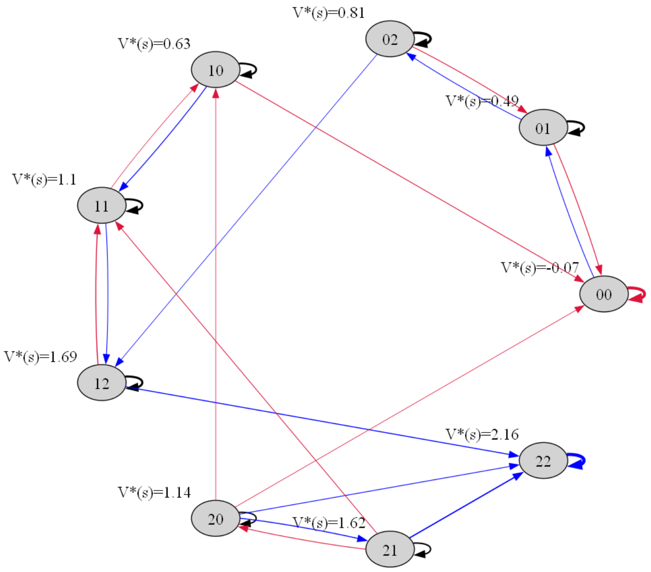


Figure 3: Transition probabilities under π^*. Only transitions with a probability of at least $\frac{1}{|S|}$ are shown and a thicker line denotes a higher probability. Blue lines are transitions to a state with a higher V^* and red lines to a state with a lower V^*.

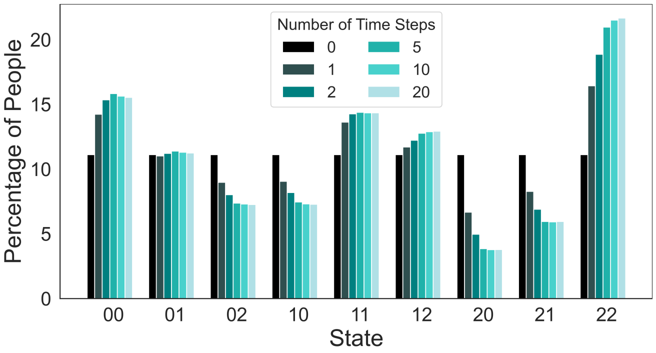


Figure 4: Percentage of people in each state after following π^* for various numbers of time steps.

exact transition probabilities can be found in Table 3 in the Appendix.

In the other states, the optimal policy π^* is to not provide human support. Figure 3 shows that in every state there is a probability of at least $\frac{1}{|S|}$ to move to a better state, with which we mean a state where V^* is higher (blue lines). Once people get to the best state 22, they have a high probability of 0.71 to stay there. However, it can also be seen that in every state except for state 22, there is a probability of at least $\frac{1}{|S|}$ to move to a worse state where V^* is lower (red lines). Especially in the worst state 00, there is a high probability of 0.64 to stay there.

Observing the effect of multiple optimal actions, Figure 4 shows that most people end up in state 22, where their appreciation for human support and self-efficacy are above average, and the expected effort and return likelihood are highest. It can also be seen that the percentage of people in state 20 is the lowest, indicating that receiving human support moves them out of this state. However, it also shows that the percentage of people in states 00, 11, and 12 increases slightly over time. In states 11 and 12, the mean reward is higher than the overall mean reward (Figure 1), so being in these states is relatively good. However, state 00 is the worst state, so we would ideally want people to move out of this state. Overall, people tend to transition to states where their behavior improves.

Q4: How do optimal and sub-optimal policies compare in their effect on users' behavior?

Setup. Now that we know the effect of the optimal policy on users' behavior, we want to compare optimal and sub-optimal policies to see whether the optimal policy actually has the best effect on how much effort users spend on their activities and on their likelihood of staying in the intervention. To do this, we compared the mean reward without cost per time step for two policies: 1) the optimal policy π^*, and 2) the worst policy π^-. We used the mean reward without cost because we only want to compare users' behavior which is described by the combined effort and return score of our reward function. The optimal policy for each state is found by taking the action which results in the highest Q-value. In our case, this results in a policy that only provides human support in state 20 (Table 1). To be able to make a good comparison, we wanted to provide a somewhat equal amount of human support for the worst policy. Therefore, we created this policy by assigning human support in the state where the difference in Q-value for providing human support and not providing human support is the smallest, i.e. the state where providing human support is the most disadvantageous. This was state 21. We simulated the different policies on the distribution of all people across the states for the first session of our study to get a general representation of people.

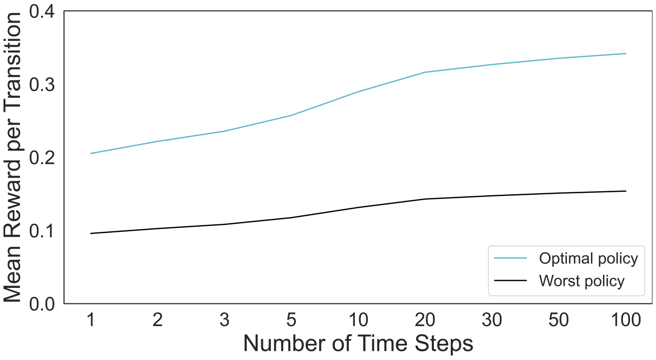


Figure 5: Mean reward per transition over time while following the optimal policy π^* and the worst policy π^-.

Results. As shown in Figure 5, the mean reward for the optimal policy π^* is highest at all time steps and increases over time, starting at 0.21 after 1 time step and ending at 0.34 after 100 time steps. This indicates that users' effort and return likelihood improve over time following the optimal policy. The mean reward for the worst policy also increases slightly, with 0.1 after 1 time step and ending at 0.15 after 100 time steps. However, the difference in mean reward increases and the optimal policy always results in a higher reward, which suggests that the optimal policy is the most successful in increasing users' effort and return likelihood over time.

Q5: How does the cost factor of a human coach in the reward function influence the reinforcement model?

Setup. Providing human support in addition to the virtual coach in a behavior change program will be more costly since the coach has to be paid. Therefore, we introduced a cost factor in our reward function to penalize the involvement of a human coach. However, this cost factor can affect the reinforcement learning model since a different reward function can lead to different optimal policies and we might want to adapt it depending on the availability of human coaches. Therefore we compared different cost factors for our reward function on four aspects: 1) the percentage of people in the best state, 2) the percentage of people in the worst state, 3) the percentage of people who received human feedback, and 4) the mean reward without cost per transition. With best and worst state we mean the states where the expected cumulative reward over time is highest and lowest, respectively. We compared these by simulating the optimal policy of the different reward functions for 20 time steps on an initial population of 500, where 10 random people left the intervention every day and 10 new people joined every day. The number of time steps was chosen, because the change in the amount of people in the best and worst states, and the percentage of people receiving human feedback stabilizes after 20 steps (see Figure 8 in the Appendix). The notion of 10 people leaving and joining each day was chosen arbitrarily to mimic the dynamics of a real-life intervention program. The states of the initial population and the states of the new people joining were based on the state distribution of the first session of our study. This simulation was run multiple times to account for the randomness of people joining/stopping. Additionally, we also analyzed the effect of following the optimal policy for varying numbers of time steps, following the approach for Q3, for a few different cost factors.

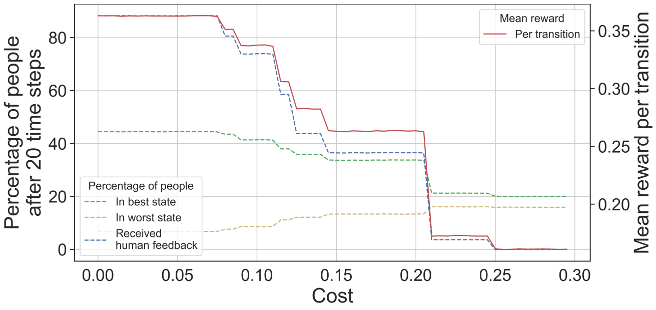
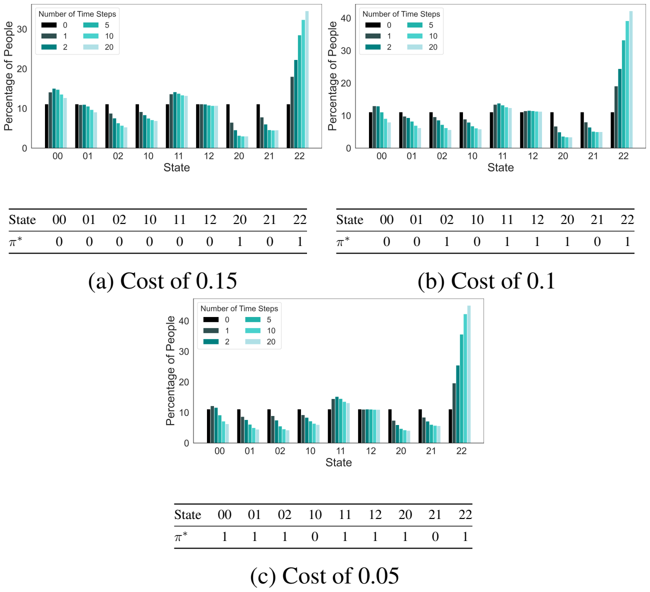


Figure 6: Right axis: Percentage of people after 20 time steps: 1) in the best state, 2) in the worst state, and 3) received human feedback for different cost factors. Left axis: Mean reward per transition after 20 time steps for different cost factors.



State	00	01	02	10	11	12	20	21	22
π^*	0	0	0	0	0	0	1	0	1

(a) Cost of 0.15

State	00	01	02	10	11	12	20	21	22
π^*	0	0	1	0	1	1	1	0	1

(b) Cost of 0.1

State	00	01	02	10	11	12	20	21	22
π^*	1	1	1	0	1	1	1	0	1

(c) Cost of 0.05

Figure 7: Optimal policy π^* and percentage of people in each state after following π^* for various numbers of time steps with different costs.

Results. Figure 6 shows that a lower cost factor in our reward function leads to a higher mean reward per transition, meaning that people are more likely to spend a high amount of effort on their activities and return to the program. This is also reflected in the fact that the percentage of people that are in the best state after 20 time steps is higher for a lower cost factor. Figure 7 shows simulations and the optimal policies for different cost factors and it can be seen that lower cost factors indeed lead to more success in moving people to the best state and away from the worst state. A larger version of this figure can be found in the Appendix (Figure 9). However, a lower cost factor will also result in a higher percentage of people receiving human support since the optimal policies will assign human support in more states. It will specifically result in providing human support in state 22, where people have a relatively high appreciation of human support and high self-efficacy. This is the best state where the expected reward is already the highest. Since the desired effect is that the amount of people in this state grows over time, it also means that the amount of people receiving human support will grow over time. This will make the intervention program more costly due to the increased need for human coaches, while the improvement in users' behavior may not increase as much since people in this state already have a relatively high effort and return likelihood. Figure 6 also shows that for lower cost factors, the ratio of the percentage of people in the best state to the amount of human support provided is much lower compared to higher cost factors above 0.21. This suggests that the benefits of human support are more pronounced at lower cost factors, as relatively much less human support is required to achieve an improvement in users' behavior. The cost factor in the reward function thus significantly influences the RL model and a higher cost factor results in a more cost-effective RL model.

4 Responsible Research

This study works with human data and implements a reinforcement learning model that can potentially be used in the health sector. Therefore, it is important to keep the ethical aspects and reproducibility of the research in mind for several reasons, such as ensuring participants' privacy, avoiding bias, enhancing transparency and maintaining scientific integrity. This section highlights the measures that have been taken to ensure this, as well as the possible limitations and recommendations.

4.1 Data collection

As mentioned in subsection 2.3, this research utilizes data from a study by Albers et al. [19]. The study was approved by the Human Research Ethics Committee of the Delft University of Technology (Letter of Approval number: 3683). Participation was voluntary and participants were asked for their informed consent multiple times during the experiments. To ensure user privacy, the data was fully anonymized before we could work with it and therefore, the participants cannot be traced back to.

To get a representative pool of participants for the study, the crowdsourcing platform Prolific was used. This also allowed for the assurance of participants who were fully vetted and verified to be real and unique people. The study aimed to get at least 40% male and at least 40% female participants, which was reached (Table 2). The distribution of the highest education level completed is also representative if we compare it to the educational attainment statistics from the US [27]. However, one downside of Prolific is that the users are paid to participate in the study. This could lead to selection bias, where we might not have a good representation of the whole population since the financial motivation could have caused people to behave differently. For example, the participants could have had more motivation to follow the smoking cessation program and therefore create a positive bias in the data. While it is difficult to fully account for this bias, we tried to mitigate some of it by analyzing the responses in relative terms. For calculating the state values, we used the percentiles, and for calculating the separate effort and return rewards we looked at the deviation from the mean of those responses. These approaches help with normalizing the data and reducing the impact of the possible bias.

4.2 Reproducibility

The reinforcement learning model that we have implemented can be reproduced by following the methods described in section 2 and the experiments to analyze the effectiveness of the model can be reproduced by following section 3. The analysis code for this research is also published on 4TU.ResearchData [28] and the data will also be published and linked to the OSF form [19]. Therefore, the results of this research can be reproduced and verified. To enhance the reproducibility, the code has been written in Jupyter notebooks and is well documented. To ensure that the code and documentation are of good quality, it has also been checked by a second person using a checklist provided by the TU Delft.

4.3 Ethical considerations

Using the RL model that we have implemented, we can determine when to provide human feedback with the limited availability of human coaches. However, this model solely bases its decisions on maximizing the effort that people spend on activities and their likelihood of staying in the program. No other principles, such as the ones mentioned by Persad [29], are considered when determining who to give human support to. However, there may be other principles, such as providing human feedback to the ones who are in the worst health conditions, that can help with determining when to involve a human coach. Thus, when integrating this RL model into an intervention program, one should also consider other principles of allocating human support to take all moral values into account.

Another ethical concern is the lack of transparency towards users of how it is decided who gets human support. Most people are unlikely to understand how reinforcement learning works and it is thus unclear to them how the model determines when to involve a human coach. This could lead to a decrease in trust in the intervention program [30]. Therefore, if this model is to be used, users should receive information and an explanation of the RL model's decision-making process to enhance their trust in the program.

5 Discussion

This research explored the use of reinforcement learning to determine when to involve a human coach in quitting smoking with a virtual coach. The results show that considering states can help with predicting the effort that people spend on preparatory activities and the likelihood of returning to the intervention (*Q1*). Furthermore, considering the current state also results in better predictions for the next states (*Q2*). This can help with choosing the action such that people are moved to future states where they are likely to spend more effort on the activities and to stay in the program. In the optimal policy, only people in state 20 receive human support. This is the state where people have a relatively high appreciation of human support and a relatively low self-efficacy. If we provide human support according to the optimal policy multiple times, people tend to move to states where their effort and return likelihood are expected to be higher. However, some people remain in state 00, where the expected effort and return likelihood is the lowest (*Q3*). The results also show that users' effort and return likelihood increase more over time when following the optimal policy compared to the worst policy (*Q4*). We also found that the cost factor in our reward function significantly impacts the RL model. A lower cost factor is more successful in moving people to better states but also results in providing significantly more human support. A higher cost factor substantially reduces the amount of human support while still improving users' behavior, making it more cost-effective (*Q5*).

Limitations and future work. While the results are promising, there are some limitations to this study. The first and foremost limitation is that the model was tested using leave-one-out cross-validation and simulations, and not on actual human subjects. This allowed us to test and fine-tune the model while staying within budget and the restricted time frame. However, we made some assumptions such as an initial population of 500 people and 10 people dropping out and joining every day for Q5, which might not be the most accurate representation of real-world scenarios. Therefore, future work should include a trial on human subjects to confirm the effectiveness of our RL model.

We also considered our transition function and reward function to be stationary as they did not change over time. However, it could be beneficial to make these dynamic to increase adaptability. For example, the cost factor could be adjusted according to the availability of human coaches. Piette et al. [17] also showed that an RL model that learns from patient interactions over time is effective and decreases the needed therapist time. For further research, one could look into dynamic transition and reward functions to further improve the model.

Next to that, our combined reward function was a simple linear combination of several objectives: effort, return likelihood, and cost factor of a human coach. We used a weighted sum for the effort and return responses, where we assumed equal importance of both objectives. This is the first limitation of our reward function because we do not know for sure if both objectives are equally important for quitting smoking. Further research could thus be done to investigate how important those factors are for quitting smoking. Secondly, using a weighted sum transformed our multi-objective problem into a single-objective one, which made it easier to work with but might not fully capture the dynamic between these objectives and oversimplify the problem. Therefore, a more complex approach to multi-objective reinforcement learning could potentially lead to better solutions. Due to the multiple objectives, there can be multiple possible optimal solutions. This set of optimal solutions is also known as the Pareto front. Instead of learning a single optimal policy, one could try to find this Pareto front by learning multiple optimal policies for different weights and costs and selecting the one that is the most suitable. This is called a multi-policy approach and there are several ways to implement this, such as Pareto Q-Learning [31] or Tree-based Fitted Q-Iteration [32]. Hayes et al. [26] also evaluated multiple multi-objective RL algorithms and encouraged the use of such algorithms for multi-objective optimization problems. Future work could investigate if such algorithms are indeed more effective for our problem and explore their use to possibly find better solutions.

Another limitation is that we only considered four possible user features for the states: appreciation of human support, self-efficacy, importance of quitting smoking, and energy level. While these features allowed us to derive efficient states for our reinforcement learning model, there could be other features that can influence a person's behavior. For example, having support from friends and family can increase the motivation to quit smoking [33]. Therefore, considering additional features in future work could help improve the accuracy of the RL model.

Lastly, the number of samples that we obtained from our study was unbalanced for each state-action pair. Table 4 in the Appendix shows the number of samples for each state-action pair and as one can see, some pairs have very few samples (e.g. 20 samples for state 21 and action 1), while others have a lot (e.g. 326 for state 00 and action 0). It makes sense that we have fewer samples for action 1 (providing human support) since a human coach was only involved for 20% of the time during the study. However, an unbalanced number of samples might lead to some skewness in our reinforcement learning model, so for further work it can be balanced by injecting some mean samples, following the approach by Dierikx [14].

6 Conclusion

The aim of this research was to analyze the effectiveness of reinforcement learning in determining when to provide human support in quitting smoking with a virtual coach. By analyzing the different components of the RL model, we have found that such a model is indeed effective and increases the effort that people spend on their activities and the likelihood of returning to the intervention. The model specifically assigns human support to people with a relatively high appreciation of human feedback and low self-efficacy. Receiving human support in this state tends to move people to a state with a higher self-efficacy, where their effort and return likelihood are higher. We also found that a higher cost factor in the reward function results in a more cost-effective allocation of human support, requiring relatively less work from human coaches to achieve an improvement in users' behavior. The insights from this study can further be used to make informed decisions about when to provide human support in smoking cessation programs.

Acknowledgements

I would like to gratefully acknowledge Nele Albers and Willem-Paul Brinkman for providing insights, feedback, and guidance throughout the process of this research. A majority of the implementation of the reinforcement learning model and the analysis steps were adapted from a similar study by Albers et al. [13]. Their work and contributions are highly appreciated. Lastly, Yoan Naydenov is thanked for checking the analysis code for this study.

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] Trimbos Instituut. "Cijfers roken." (2023), [Online]. Available: https://www.trimbos.nl/kennis/cijfers/roken/.

[2] World Health Organization. "Tobacco." (), [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco.

[3] M. Chaiton, L. Diemert, J. E. Cohen, *et al.*, "Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers," en, *BMJ Open*, vol. 6, no. 6, e011045, Jun. 2016, ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2016-011045. [Online]. Available: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2016-011045 (visited on 05/08/2024).

[4] S.-H. Zhu, T. Melcer, J. Sun, B. Rosbrook, and J. P. Pierce, "Smoking cessation with and without assistance," en, *American Journal of Preventive Medicine*, vol. 18, no. 4, pp. 305–311, May 2000, ISSN: 07493797. DOI: 10.1016/S0749-3797(00)00124-0. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0749379700001240 (visited on 05/08/2024).

[5] Y. Duan, B. Shang, W. Liang, G. Du, M. Yang, and R. E. Rhodes, "Effects of eHealth-based multiple health behavior change interventions on physical activity, healthy diet, and weight in people with noncommunicable diseases: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 23, no. 2, e23786, Feb. 22, 2021, ISSN: 1439-4456. DOI: 10.2196/23786. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8074786/ (visited on 05/22/2024).

[6] K. Gemesi, S. Winkler, S. Schmidt-Tesch, F. Schederecker, H. Hauner, and C. Holzapfel, "Efficacy of an app-based multimodal lifestyle intervention on body weight in persons with obesity: Results from a randomized controlled trial," *International Journal of Obesity*, vol. 48, no. 1, pp. 118–126, Jan. 2024, Publisher: Nature Publishing Group, ISSN: 1476-5497. DOI: 10.1038/s41366-023-01415-0. [Online]. Available: https://www.nature.com/articles/s41366-023-01415-0 (visited on 05/22/2024).

[7] P. Chikersal, D. Belgrave, G. Doherty, *et al.*, "Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention," en, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–16, ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376341. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376341 (visited on 04/23/2024).

[8] Y.-C. Lee, N. Yamashita, and Y. Huang, "Exploring the Effects of Incorporating Human Experts to Deliver Journaling Guidance through a Chatbot," en, *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–27, Apr. 2021, ISSN: 2573-0142. DOI: 10.1145/3449196. [Online]. Available: https://dl.acm.org/doi/10.1145/3449196 (visited on 04/23/2024).

[9] N. Albers, M. A. Neerincx, K. M. Penfornis, and W.-P. Brinkman, "Users' needs for a digital smoking cessation application and how to address them: A mixed-methods study," *PeerJ*, vol. 10, e13824, Aug. 19, 2022, ISSN: 2167-8359. DOI: 10.7717/peerj.13824. [Online]. Available: https://peerj.com/articles/13824 (visited on 06/18/2024).

[10] D. Mohr, P. Cuijpers, and K. Lehman, "Supportive accountability: A model for providing human support to enhance adherence to eHealth interventions," *Journal of Medical Internet Research*, vol. 13, no. 1, e1602, Mar. 10, 2011, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. DOI: 10.2196/jmir.1602. [Online]. Available: https://www.jmir.org/2011/1/e30 (visited on 05/22/2024).

[11] S. Michie, M. M. Van Stralen, and R. West, "The behaviour change wheel: A new method for characterising and designing behaviour change interventions," *Implementation Science*, vol. 6, no. 1, p. 42, Dec. 2011, ISSN: 1748-5908. DOI: 10.1186/1748-5908-6-42. [Online]. Available: http://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-6-42 (visited on 06/03/2024).

[12] N. Albers, M. A. Neerincx, and W.-P. Brinkman, "Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active," en, *PLOS ONE*, vol. 17, no. 12, F. L. Wang, Ed., e0277295, Dec. 2022, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0277295. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0277295 (visited on 04/23/2024).

[13] 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," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '23, Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2023, pp. 717–726, ISBN: 9781450394321.

[14] 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*, N. Baghaei, R. Ali, K. Win, and K. Oyibo, Eds., Cham: Springer Nature Switzerland, 2024, pp. 100–115, ISBN: 978-3-031-58226-4. DOI: 10.1007/978-3-031-58226-4_9.

[15] T. Weimann and C. Gißke, "Unleashing the Potential of Reinforcement Learning for Personalizing Behavioral Transformations with Digital Therapeutics: A Systematic Literature Review:" en, 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, ISBN: 978-989-758-688-0. DOI: 10.5220/0012474700003657. [Online]. Available: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0012474700003657 (visited on 05/02/2024).

[16] J. D. Piette, S. Newman, S. L. Krein, *et al.*, "Patient-centered pain care using artificial intelligence and mobile health tools: A randomized comparative effectiveness trial," *JAMA Internal Medicine*, vol. 182, no. 9, pp. 975–983, Sep. 1, 2022, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2022.3178. [Online]. Available: https://doi.org/10.1001/jamainternmed.2022.3178 (visited on 05/29/2024).

[17] J. D. Piette, S. Newman, S. L. Krein, *et al.*, "Artificial intelligence (AI) to improve chronic pain care: Evidence of AI learning," *Intelligence-Based Medicine*, vol. 6, p. 100 064, Jan. 1, 2022, ISSN: 2666-5212. DOI: 10.1016/j.ibmed.2022.100064. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666521222000175 (visited on 06/05/2024).

[18] E. M. Forman, S. G. Kerrigan, M. L. Butryn, *et al.*, "Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss?" en, *Journal of Behavioral Medicine*, vol. 42, no. 2, pp. 276–290, Apr. 2019, ISSN: 0160-7715, 1573-3521. DOI: 10.1007/s10865-018-9964-1. [Online]. Available: http://link.springer.com/10.1007/s10865-018-9964-1 (visited on 05/02/2024).

[19] N. Albers and W.-P. Brinkman, "Perfect fit - learning when to involve a human coach in an ehealth application for preparing for quitting smoking or vaping," 2024. DOI: https://doi.org/10.17605/OSF.IO/78CNR. [Online]. Available: https://osf.io/78cnr.

[20] N. Albers, *Virtual coach kai for preparing for quitting smoking with human support*, May 2024. [Online]. Available: https://doi.org/10.5281/zenodo.11102861.

[21] R. P. Ghantasala, N. Albers, K. M. Penfornis, M. H. M. van Vliet, and W.-P. Brinkman, "Feasibility of generating structured motivational messages for tailored physical activity coaching," *Frontiers in Digital Health*, vol. 5, Sep. 12, 2023, Publisher: Frontiers, ISSN: 2673-253X. DOI: 10.3389/fdgth.2023.1215187. [Online]. Available: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1215187/full (visited on 06/03/2024).

[22] E. McAuley, C. Lox, and T. E. Duncan, "Long-term maintenance of exercise, self-efficacy, and physiological change in older adults," *Journal of Gerontology*, vol. 48, no. 4, P218–P224, Jul. 1, 1993, ISSN: 0022-1422. DOI: 10.1093/geronj/48.4.P218. [Online]. Available: https://academic.oup.com/geronj/article-lookup/doi/10.1093/geronj/48.4.P218 (visited on 05/16/2024).

[23] N. B. Rajani, N. Mastellos, and F. T. Filippidis, "Self-efficacy and motivation to quit of smokers seeking to quit: Quantitative assessment of smoking cessation mobile apps," *JMIR Mhealth Uhealth*, vol. 9, no. 4, e25030, Apr. 2021, ISSN: 2291-5222. DOI: https://doi.org/10.2196/25030. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33929336.

[24] D. Chapman and L. P. Kaelbling, "Input generalization in delayed reinforcement learning: An algorithm and performance comparisons," in *International Joint Conference on Artificial Intelligence*, 1991. [Online]. Available: https://api.semanticscholar.org/CorpusID:7213327.

[25] J. C. Hutchinson and G. Tenenbaum, "Perceived effort — can it be considered gestalt?" *Psychology of Sport and Exercise*, vol. 7, no. 5, pp. 463–476, 2006, ISSN: 1469-0292. DOI: https://doi.org/10.1016/j.psychsport.2006.01.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1469029206000100.

[26] C. F. Hayes, R. Rădulescu, E. Bargiacchi, *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, p. 26, Apr. 13, 2022, ISSN: 1573-7454. DOI: 10.1007/s10458-022-09552-y. [Online]. Available: https://doi.org/10.1007/s10458-022-09552-y (visited on 06/18/2024).

[27] Census Bureau. "Census bureau releases new educational attainment data." (Feb. 16, 2023), [Online]. Available: https://www.census.gov/newsroom/press-releases/2023/educational-attainment-data.html.

[28] S. Li, *Analysis code for bachelor thesis: Using reinforcement learning to determine when to provide human support in quitting smoking with a virtual coach*, Jun. 2024. DOI: 10.4121/19dc8011-1bcb-4143-a373-08718055dc7c. [Online]. Available: https://data.4tu.nl/datasets/19dc8011-1bcb-4143-a373-08718055dc7c.

[29] G. Persad, A. Wertheimer, and E. J. Emanuel, "Principles for allocation of scarce medical interventions," *The Lancet*, vol. 373, no. 9661, pp. 423–431, Jan. 2009, ISSN: 01406736. DOI: 10.1016/S0140-6736(09)60137-9. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0140673609601379 (visited on 06/18/2024).

[30] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust AI," *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, Dec. 1, 2021, ISSN: 2210-5441. DOI: 10.1007/s13347-021-00477-0. [Online]. Available: https://doi.org/10.1007/s13347-021-00477-0 (visited on 06/20/2024).

[31] K. V. Moffaert and A. Nowe, "Multi-objective reinforcement learning using sets of pareto dominating policies," *The Journal of Machine Learning Research*, vol. 15, pp. 3483–3512, Nov. 2014.

[32] A. Castelletti, F. Pianosi, and M. Restelli, "Tree-based fitted q-iteration for multi-objective markov decision problems," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Brisbane, Australia: IEEE, Jun. 2012, pp. 1–8. DOI: 10.1109/IJCNN.2012.6252759. [Online]. Available: http://ieeexplore.ieee.org/document/6252759/ (visited on 06/20/2024).

[33] J. N. Soulakova, C.-Y. Tang, S. A. Leonardo, and L. A. Taliaferro, "Motivational benefits of social support and behavioural interventions for smoking ces-

sation," *Journal of smoking cessation*, vol. 13, no. 4, pp. 216–226, Dec. 2018, ISSN: 1834-2612. DOI: 10.1017/jsc.2017.26. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6459678/ (visited on 06/05/2024).

A Participant Characteristics

Table 2: Characteristics of the 678 people with at least one transition sample

Characteristic	**Value**
Age	
- 18 – 20, n (%)	12 (1.77%)
- 21 – 30, n (%)	235 (34.66%)
- 31 – 40, n (%)	213 (31.42%)
- 41 – 50, n (%)	128 (18.88%)
- 51 – 60, n (%)	73 (10.77%)
- 61 – 70, n (%)	16 (2.36%)
- 71 – 80, n (%)	1 (0.15%)
Gender	
- Man (including Trans Male/Trans Man), n (%)	334 (49.26%)
- Woman (including Trans Female/Trans Woman), n (%)	330 (48.67%)
- Non-binary (would like to give more detail), n (%)	14 (2.06%)
- Rather not say, n (%)	0 (0.0%)
Highest education level completed	
- No formal qualifications, n (%)	5 (0.74%)
- Secondary education (e.g. GED/GCSE), n (%)	61 (9.0%)
- High school diploma/A-levels, n (%)	139 (20.5%)
- Technical/community college, n (%)	89 (13.13%)
- Undergraduate degree (BA/BSc/other), n (%)	263 (38.79%)
- Graduate degree (MA/MSc/MPhil/other), n (%)	107 (15.78%)
- Doctorate degree (PhD/other), n (%)	9 (1.33%)
- Don't know / not applicable, n (%)	5 (0.74%)

B Transition probabilities

Table 3: Transition probabilities $T^*(s, s')$ from state s to state s' following optimal policy π^*. Only probabilities higher than $\frac{1}{|S|}$ are shown.

s	s'	$T^*(s, s')$	s	s'	$T^*(s, s')$	s	s'	$T^*(s, s')$
00	00	0.64	10	00	0.17	20	00	0.12
00	01	0.16	10	10	0.42	20	10	0.12
01	00	0.18	10	11	0.19	20	20	0.29
01	01	0.47	11	10	0.12	20	21	0.21
01	02	0.16	11	11	0.44	20	22	0.17
02	01	0.17	11	12	0.16	21	11	0.16
02	02	0.55	12	11	0.21	21	20	0.17
02	12	0.12	12	12	0.47	21	21	0.31
			12	22	0.21	21	22	0.27
						22	22	0.71

C Simulation of the RL model

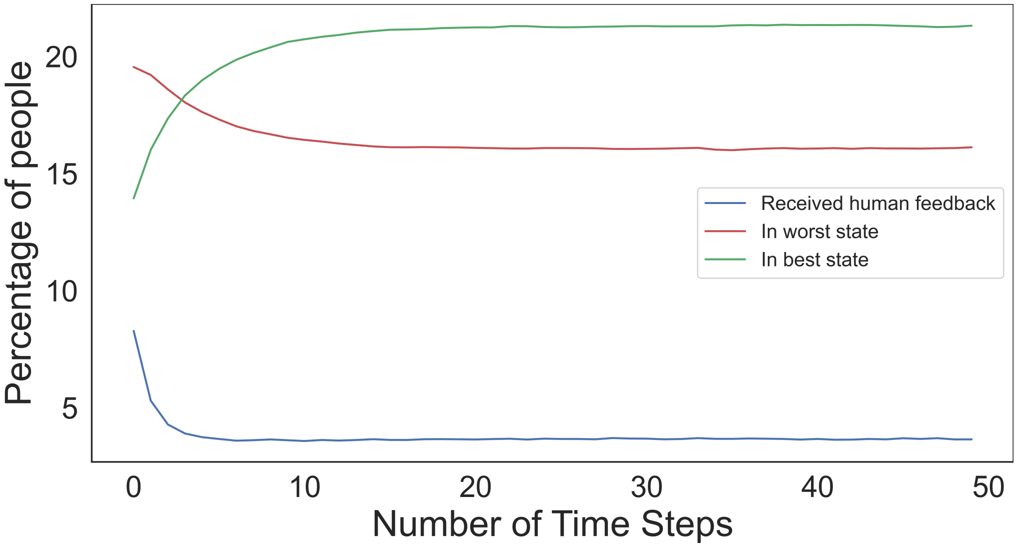


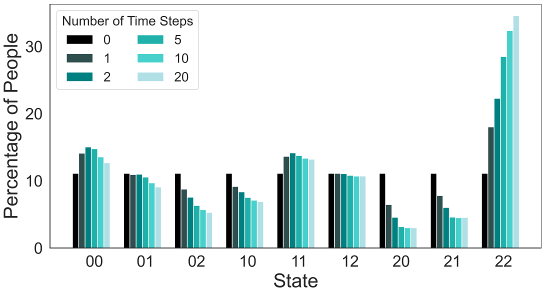
Figure 8: Percentage of people 1) who received human feedback, 2) in the worst state, and 3) in the best state following π^*.

D Sample sizes

Table 4: Number of samples for each state-action pair

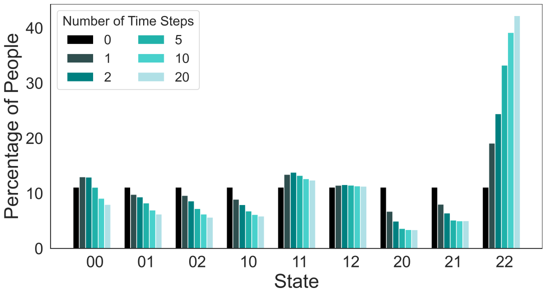
State	Action	Number of samples
00	0	326
00	1	86
01	0	216
01	1	51
02	0	139
02	1	32
10	0	194
10	1	56
11	0	295
11	1	76
12	0	191
12	1	44
20	0	121
20	1	24
21	0	118
21	1	20
22	0	259
22	1	75

E Simulations and optimal policies for different costs



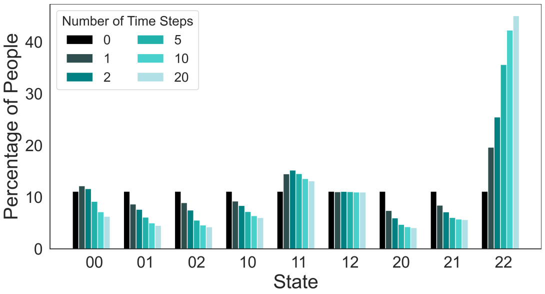
State	00	01	02	10	11	12	20	21	22
π^*	0	0	0	0	0	0	1	0	1

(a) Cost of 0.15



State	00	01	02	10	11	12	20	21	22
π^*	0	0	1	0	1	1	1	0	1

(b) Cost of 0.1



State	00	01	02	10	11	12	20	21	22
π^*	1	1	1	0	1	1	1	0	1

(c) Cost of 0.05

Figure 9: Optimal policy π^* and percentage of people in each state after following π^* for various numbers of time steps with different costs.