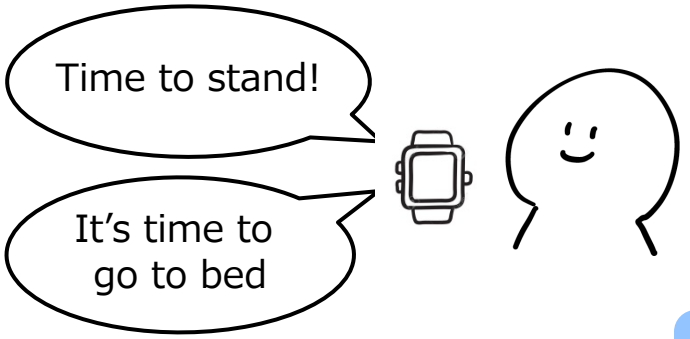
Can Reinforcement Learning Lead to Healthy Life?: Simulation Study Based on User Activity Logs

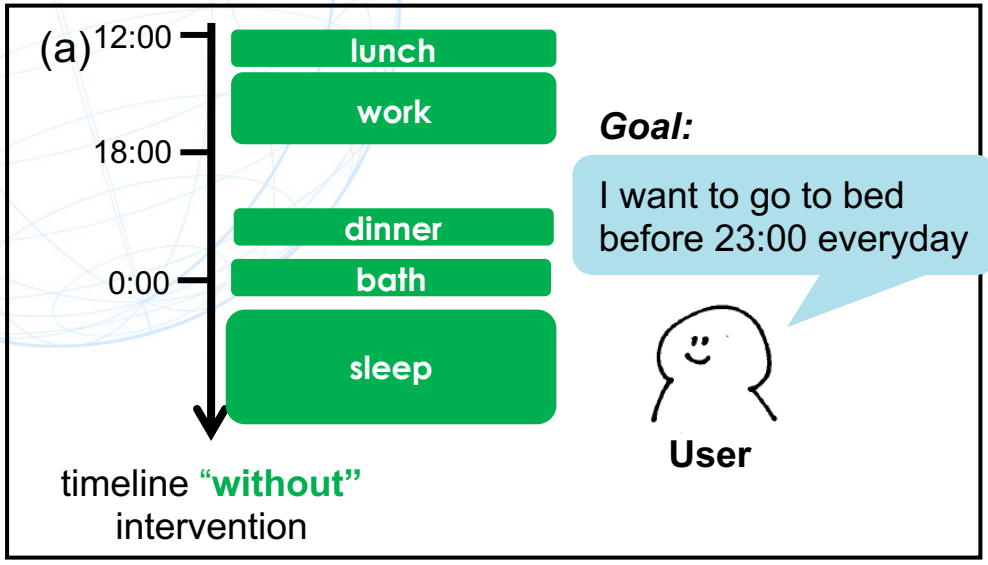
Masami Takahashi, Masahiro Kohjima, Takeshi Kurashima,
Hiroyuki Toda (NTT Service Evolution Laboratories)

Introduction

- Recently, the widespread use of applications and devices (e.g., Fitbit, Apple watch) that promote healthy behaviors has made it easier to collect health data and intervene such as notifying the user of health-related information.
- However, the interventions realized by these applications are too simple. It seems that current apps provide many *ineffective* interventions that the user is unlikely to follow.

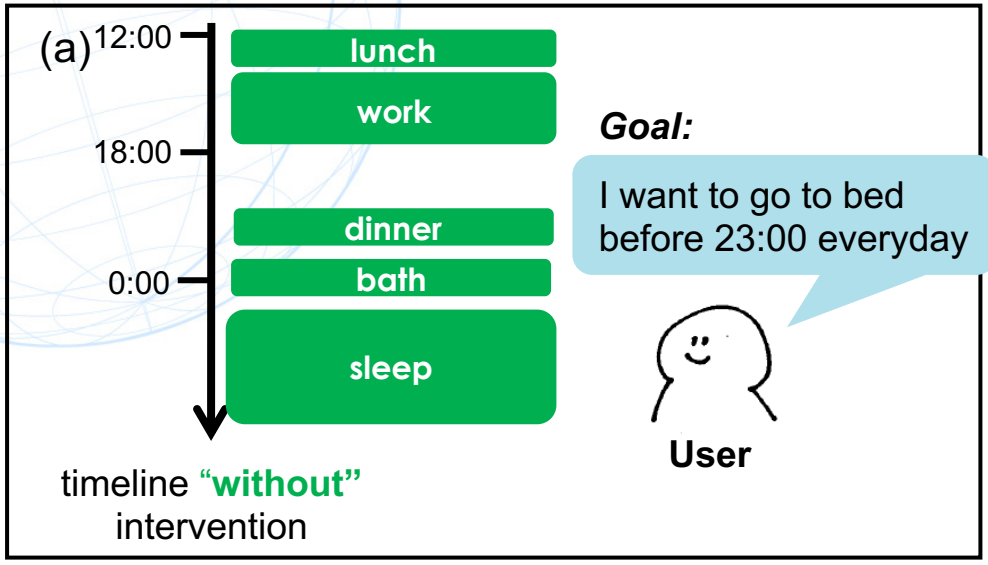


Example

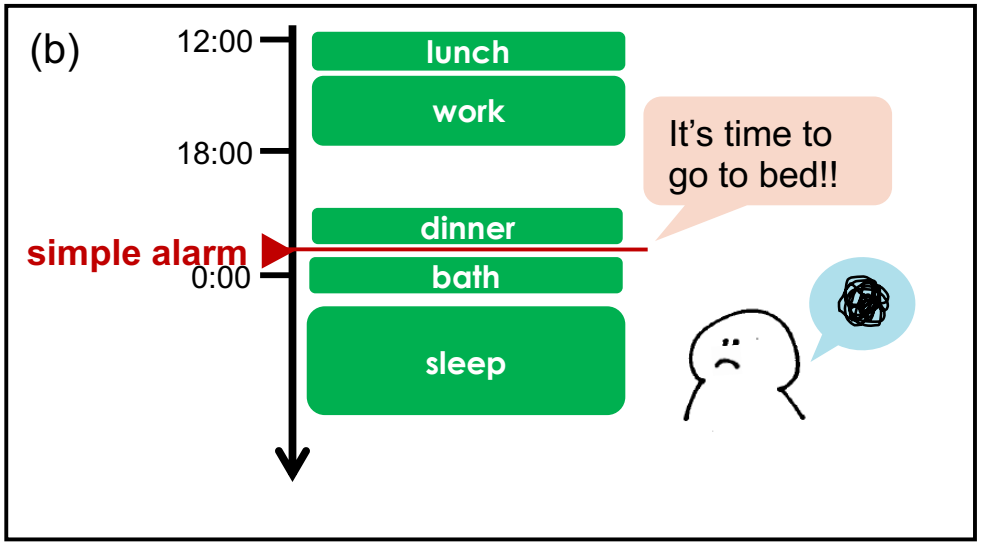


(a) The user specifies the goal (sleep at 11:00 p.m.).

Example

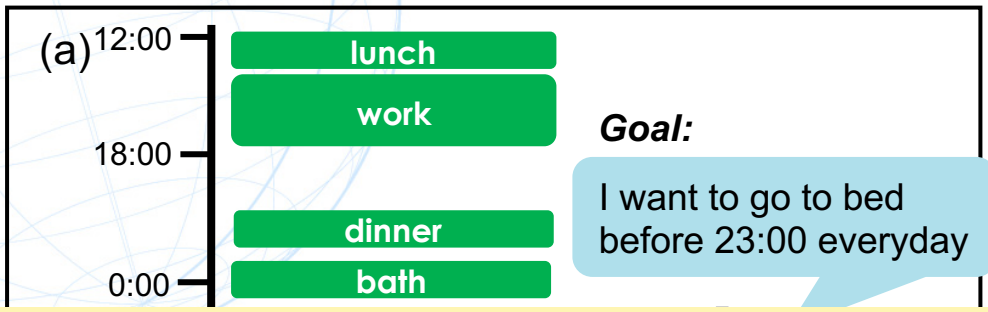


(a) The user specifies the goal (sleep at 11:00 p.m.).



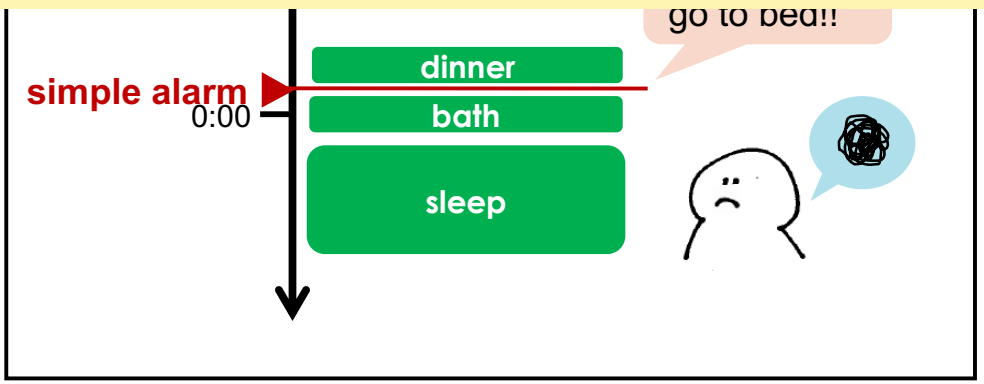
(b) It is difficult to respond to a simple intervention if the activities that must be performed prior to sleeping have not been completed.

Example



(a) The user specifies the goal (sleep at 11:00 p.m.).

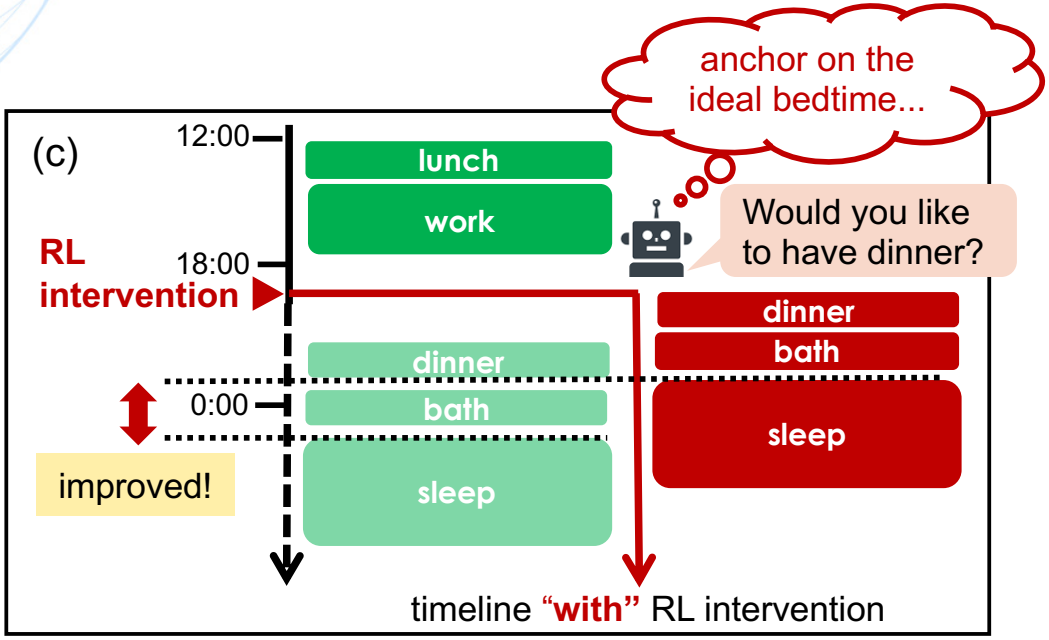
It is desirable to encourage not only the desired activity but also prior activities given their effect on the user's decisions.



the activities that must be performed prior to sleeping have not been completed.

Key idea: planning backward from the goals

- Our agent encourages the user to complete prior dependent actions with the goal of sleep.



- By encouraging the user to take dinner around 6:00 p.m., it changes the user's future state (taking bath earlier than usual) and makes it possible for the user reach the goal.

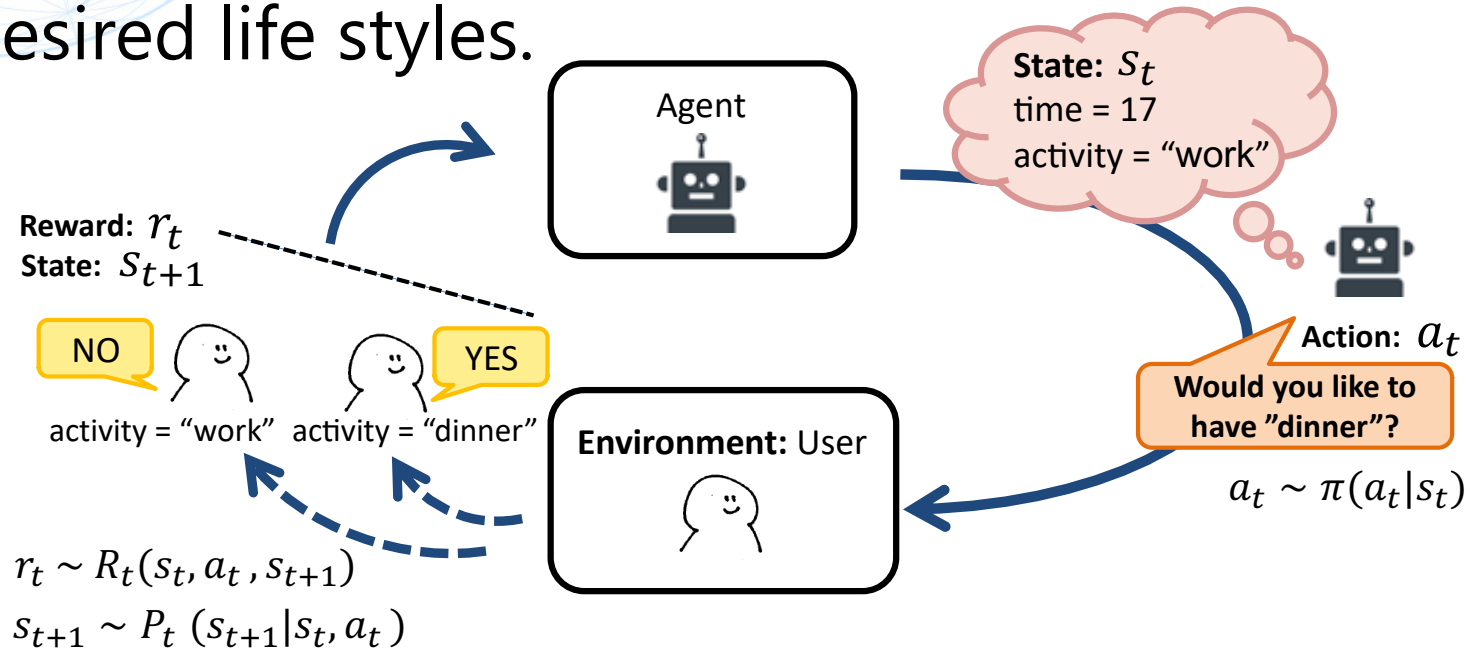
Our Approach: Reinforcement Learning (1/2)

- Considering the user's health goal, application should provide intervention at the appropriate timing to help the user achieve the goal.

- The reinforcement learning (RL) approach is well suited to this type of problem since RL makes decisions based on planning that consider the effect of a current decision on the future.

Our Approach: Reinforcement Learning (2/2)

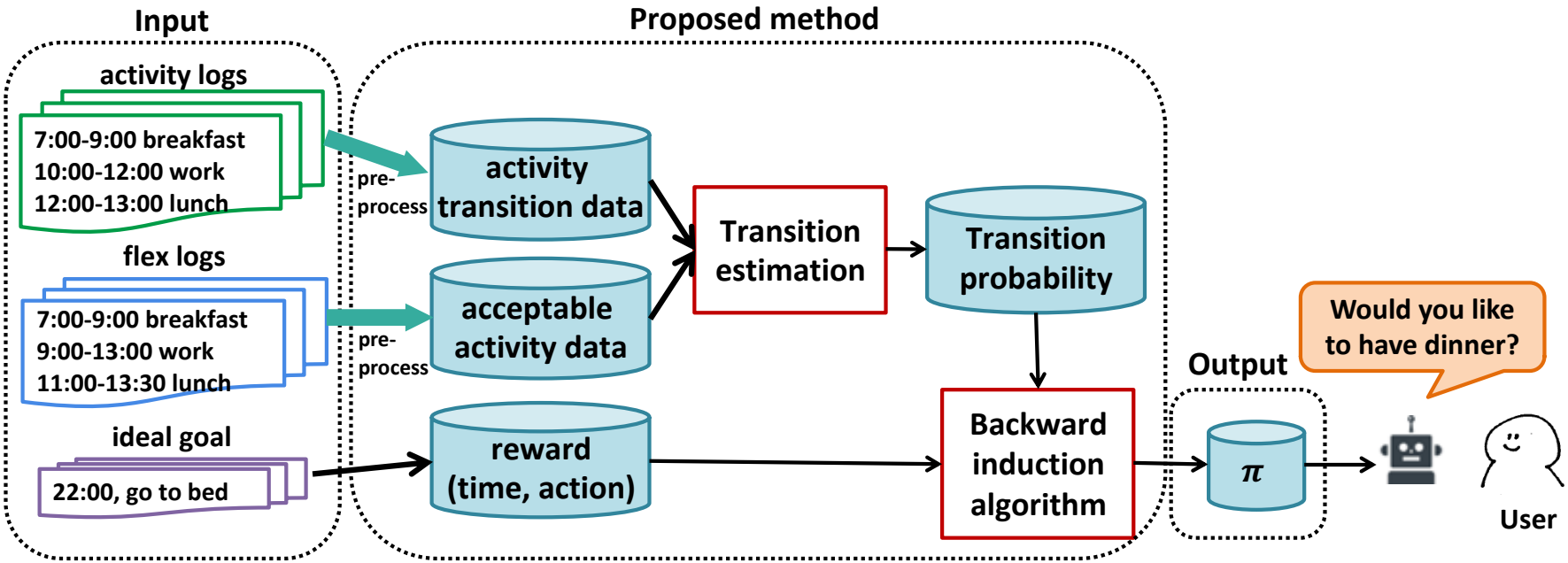
- We propose an automatic intervention method based on RL and investigate the effects of RL-based intervention to help users achieve their desired life styles.



- State, action and reward in this RL problem correspond to the user's activity, app's intervention, and user's goal.

Overview of the Proposed Method

- Input: activity logs, flex logs, user's goal (reward)
- We estimate transition probability from the logs. The optimal policy is generated by backward induction algorithm.



Transition probability estimation

- Model of the transition probability:

$$P_t^{\theta}(s_{t+1} = j | s_t = i, a = k) \text{ to } P_{tijk}^{\theta}$$

- Likelihood function of the activity transition data:

$$P(\mathcal{D}^{tr}|\theta) = \prod_{t=1}^{T} \prod_{i,j \in \mathcal{S}} \left(p_{tij|\mathcal{A}|}^{\theta}\right)^{N_{tij}}$$

- Likelihood function of the acceptable activity data:

$$P(\mathcal{D}^{apt}|\theta) = \prod_{t=1}^{T} \prod_{i,j \in \mathcal{S}} \left(p_{tijj}^{\theta}\right)^{\beta M_{tj}} \left(1 - p_{tijj}^{\theta}\right)^{\{(1-\beta)M_{tj} + (L - M_{tj})\}}$$

- Objective function:

$$\mathcal{L}(\theta) = -\log P(\mathcal{D}^{tr}|\theta) - \gamma \log P(\mathcal{D}^{apt}|\theta) + \Omega(\theta),$$

- θ is estimated by optimizing this objective:

$$\hat{\theta} = \arg min_{\theta}\, \mathcal{L}(\theta)$$

Backward Induction Algorithm

- Given the estimated transition probability and reward function, our system outputs the optimal policy by value iteration.

Algorithm 1 Backward Induction Algorithm for Finite-Horizon Entropy-regularized RL

Input: $\mathcal{P}$: transition probability, $\mathcal{R}$: reward function, α: hyperparameter

Output: $\{Q_t^*\}_t, \{V_t^*\}_t$: value function, $\{\pi_t^*\}_t$: policy

1: Set $t \leftarrow T$ and $V_T(s) = 0$ for all $s \in \mathcal{S}$.

2: Set $t \leftarrow t - 1$

3: Compute $Q_t(s, a)$ following

$$Q_t^{\pi^*}(s, a) = \mathbb{E}_{s' \sim \mathcal{P}_t(s'|s,a)}[\mathcal{R}_t(s, a, s') + V_{t+1}^{\pi^*}(s')]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

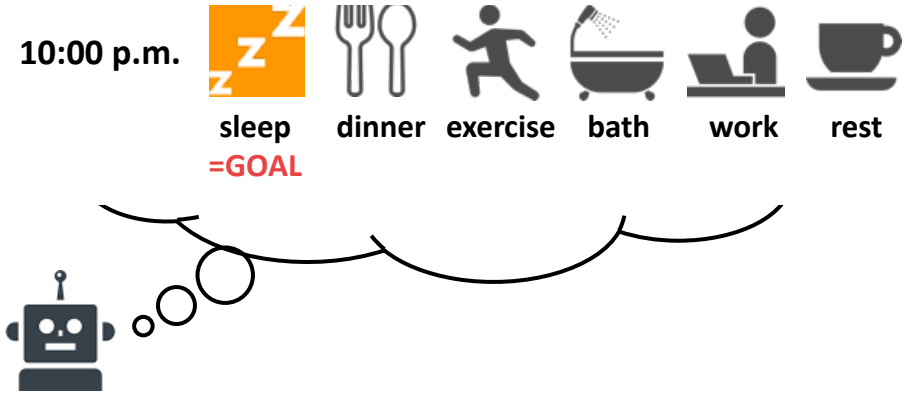
4: Compute $V_t(s)$ for all $s \in \mathcal{S}$ following

$$V_t^{\pi^*}(s) = \alpha \log \sum_{a'} \exp\left(\alpha^{-1} Q_t^{\pi^*}(s, a')\right).$$

5: Compute $\pi_t(a|s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ following

$$\pi_t^*(a|s) = \exp\left(\alpha^{-1}\{Q_t^{\pi^*}(s, a) - V_t^{\pi^*}(s)\}\right).$$

6: If $t = 0$, stop. Otherwise, return to step 2.



Backward Induction Algorithm

- Given the estimated transition probability and reward function, our system outputs the optimal policy by value iteration.

Algorithm 1 Backward Induction Algorithm for Finite-Horizon Entropy-regularized RL

Input: $\mathcal{P}$: transition probability, $\mathcal{R}$: reward function, α: hyperparameter

Output: $\{Q_t^*\}_t, \{V_t^*\}_t$: value function, $\{\pi_t^*\}_t$: policy

1: Set $t \leftarrow T$ and $V_T(s) = 0$ for all $s \in \mathcal{S}$.

2: Set $t \leftarrow t - 1$

3: Compute $Q_t(s, a)$ following

$$Q_t^{\pi^*}(s,a) = \mathbb{E}_{s' \sim \mathcal{P}_t(s'|s,a)}[\mathcal{R}_t(s,a,s') + V_{t+1}^{\pi^*}(s')]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

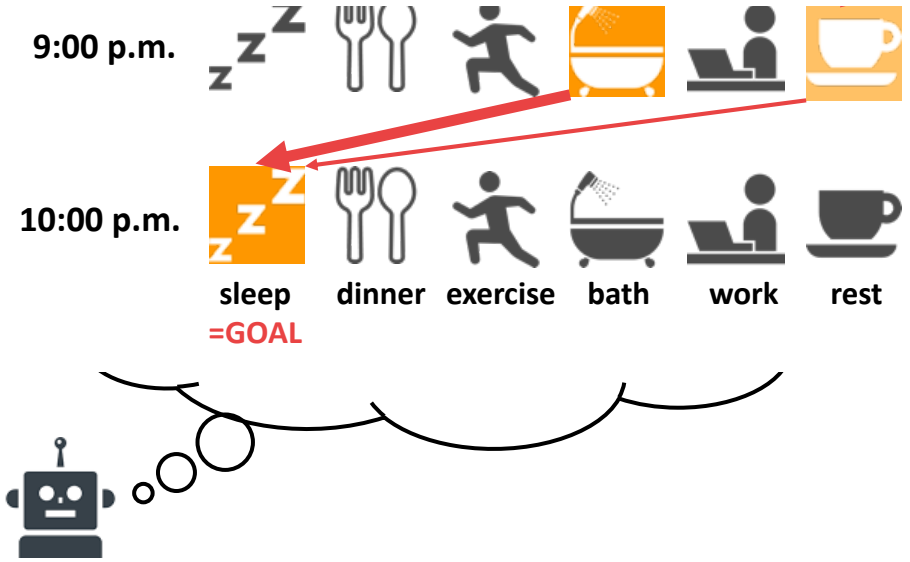
4: Compute $V_t(s)$ for all $s \in \mathcal{S}$ following

$$V_t^{\pi^*}(s) = \alpha \log \sum\nolimits_{a'} \exp\left(\alpha^{-1} Q_t^{\pi^*}(s,a')\right).$$

5: Compute $\pi_t(a|s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ following

$$\pi_t^*(a|s) = \exp\left(\alpha^{-1}\{Q_t^{\pi^*}(s,a) - V_t^{\pi^*}(s)\}\right).$$

6: If $t = 0$, stop. Otherwise, return to step 2.



Backward Induction Algorithm

- Given the estimated transition probability and reward function, our system outputs the optimal policy by value iteration.

Algorithm 1 Backward Induction Algorithm for Finite-Horizon Entropy-regularized RL

Input: $\mathcal{P}$: transition probability, $\mathcal{R}$: reward function, α: hyperparameter

Output: $\{Q_t^*\}_t, \{V_t^*\}_t$: value function, $\{\pi_t^*\}_t$: policy

1: Set $t \leftarrow T$ and $V_T(s) = 0$ for all $s \in \mathcal{S}$.

2: Set $t \leftarrow t - 1$

3: Compute $Q_t(s, a)$ following

$$Q_t^{\pi^*}(s,a) = \mathbb{E}_{s' \sim \mathcal{P}_t(s'|s,a)}[\mathcal{R}_t(s,a,s') + V_{t+1}^{\pi^*}(s')]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

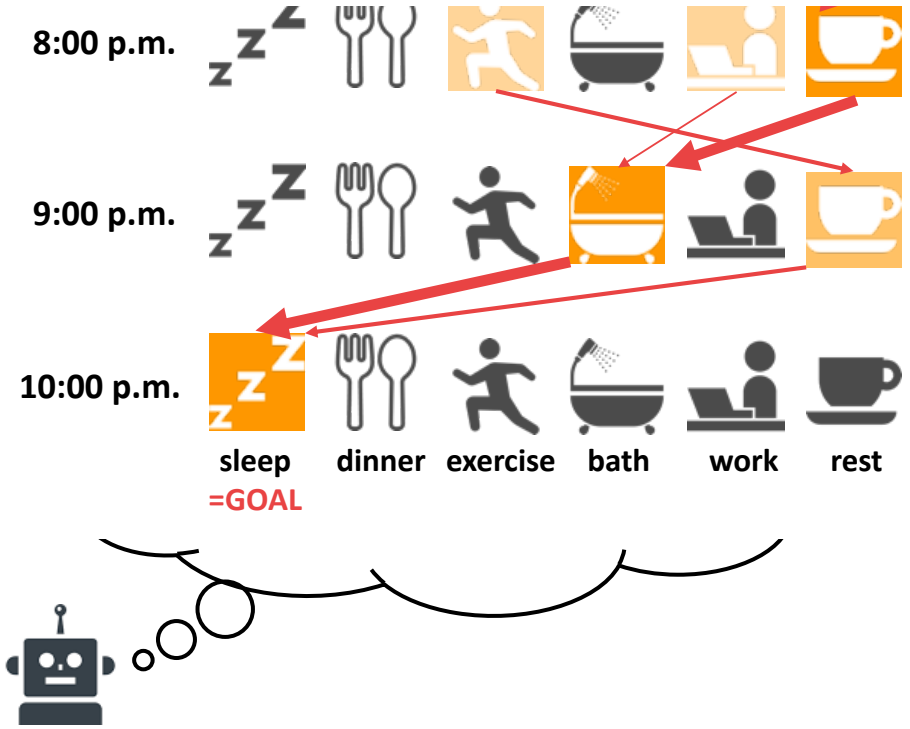
4: Compute $V_t(s)$ for all $s \in \mathcal{S}$ following

$$V_t^{\pi^*}(s) = \alpha \log \sum\nolimits_{a'} \exp\left(\alpha^{-1} Q_t^{\pi^*}(s,a')\right).$$

5: Compute $\pi_t(a|s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ following

$$\pi_t^*(a|s) = \exp\left(\alpha^{-1}\{Q_t^{\pi^*}(s,a) - V_t^{\pi^*}(s)\}\right).$$

6: If $t = 0$, stop. Otherwise, return to step 2.



Backward Induction Algorithm

- Given the estimated transition probability and reward function, our system outputs the optimal policy by value iteration.

Algorithm 1 Backward Induction Algorithm for Finite-Horizon Entropy-regularized RL

Input: $\mathcal{P}$: transition probability, $\mathcal{R}$: reward function, α: hyperparameter

Output: $\{Q_t^*\}_t, \{V_t^*\}_t$: value function, $\{\pi_t^*\}_t$: policy

1: Set $t \leftarrow T$ and $V_T(s) = 0$ for all $s \in \mathcal{S}$.

2: Set $t \leftarrow t - 1$

3: Compute $Q_t(s, a)$ following

$$Q_t^{\pi^*}(s,a) = \mathbb{E}_{s' \sim \mathcal{P}_t(s'|s,a)}[\mathcal{R}_t(s,a,s') + V_{t+1}^{\pi^*}(s')]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

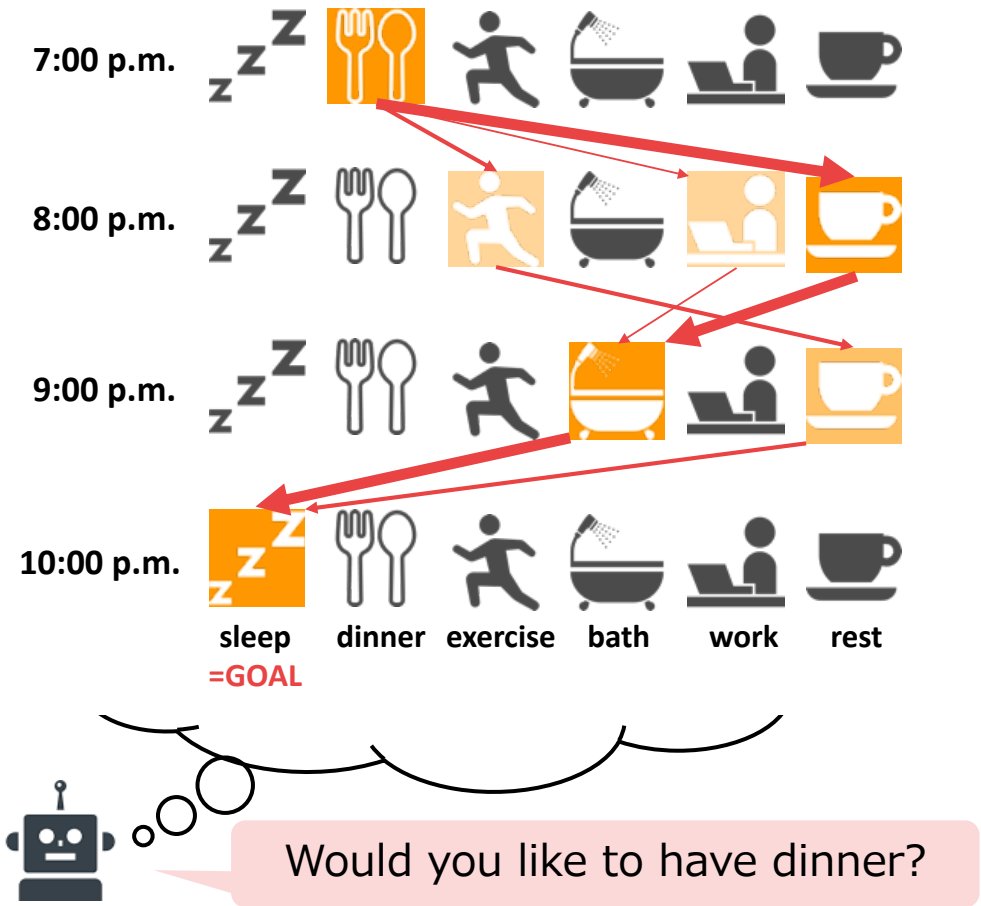
4: Compute $V_t(s)$ for all $s \in \mathcal{S}$ following

$$V_t^{\pi^*}(s) = \alpha \log \sum_{a'} \exp\left(\alpha^{-1} Q_t^{\pi^*}(s,a')\right).$$

5: Compute $\pi_t(a|s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ following

$$\pi_t^*(a|s) = \exp\left(\alpha^{-1}\{Q_t^{\pi^*}(s,a) - V_t^{\pi^*}(s)\}\right).$$

6: If $t = 0$, stop. Otherwise, return to step 2.



Experiment: Simulated Interventions

- By collecting real user data from 34 participants over a 2-month period, we construct a user simulation model.
- Reward: “good sleep reward”
- Goal time for each participant:
  - (a) mode time
  - (b) mode time -1 hours
  - (c) mode time -2 hours
- Baseline policies:
  - random intervention
  - alarm settings (one time reminder)
- Performance metric: average return

Result

- We compared the average return by the proposed method with the baselines.
- The results show our method attained the highest rewards.

TABLE II: Average reward values of all participants (proposed method, random, one time). Larger is better.

	proposed	random	one time
mode	59.23(±19.38)	-22.17(±23.51)	55.70(±20.02)
mode-1	23.07(±23.54)	-53.47(±19.62)	20.60(±22.23)
mode-2	12.77(±20.14)	-58.88(±15.32)	11.65(±19.30)

Conclusion

- We consider automatic intervention methods based on RL for lifestyle improvement.
- By collecting real user data from 34 participants over a 2-month period, we construct a user simulation model.
- We conduct simulations to evaluate the effectiveness of the RL method. The results show that the proposed method can output interventions that are more effective for goal achievement than random intervention or simple alarm setting.