IEEE Robotics and Automation Letters. This is the author's version of an article that has been published. Changes were made to this version by the publisher prior to publication. (https://doi.org/10.1109/LRA.2022.3140813)

Hybrid Hierarchical Learning for Adaptive Persuasion in Human-Robot Interaction

Shane Saunderson, *Student Member, IEEE* and Goldie Nejat, *Member, IEEE*

***Abstract*—Adaptive learning is critical to helping robots personalize their interactions with people, particularly when considering skills needed by socially assistive robots, such as persuasion. In this paper, we propose a novel, hybrid hierarchical learning architecture for use in social human-robot interaction (HRI) to adapt robot persuasive behaviors to both the static (e.g. need for cognition) and dynamic (e.g. affect) considerations of a user. A learning hierarchy is introduced that uses a contextual bandit approach in the top level to optimize for a static cognition bias and *Q*-Learning in the lower level to optimize selection of a robot persuasive strategy to deploy that aligns with a user's affect. We compare the performance of our system with a non-hierarchical learning method in simulated experiments for the task of persuading people to do daily exercises. The results show that our hybrid hierarchical architecture outperforms a non-hierarchical benchmark in learning speed and robustness to both longitudinal user change and noisy observations. Our architecture is the first to: 1) persuasively adapt to different users during social HRI considering both static and dynamic user change, and 2) use user state decomposition in persuasive HRI.**

***Index Terms*—Social HRI, Human-Centered Robotics, Reinforcement Learning, Robot Persuasion, Hybrid Architectures**

I. Introduction

As assistive robots continue to be deployed into care environments such as hospitals and assisted living facilities [1], they encounter new challenges requiring new skills. While the care tasks robots need to complete often have physical aspects to them, many are social in context. These tasks can require the use of social skills such as motivation [2] and persuasion [3]. Persuasion can be utilized by assistive robots to motivate exercise and physiotherapy [4], promote medication adherence [5], and much more.

The challenge with using persuasion for these types of social tasks is that a socially assistive robot is often required to learn and adapt to individual users, their preferences, and their abilities. This necessitates the incorporation of adaptive systems for such robots. In particular, *Adaptive Persuasive Systems* (APS) can be used, which are systems that acquire user information, develop or update user models, and adapt their persuasive approaches to personalize to an individual [6]. APS have been shown to be more effective in persuasive tasks than non-adaptive systems [7]. To date, they have been utilized in social media [8], emails [9], and text messaging [10].

Limited research has been conducted on the use of APS by social robots. In general, robots leverage adaptive systems during human-robot interaction (HRI) that are either [11]: 1) *reactive*, that adapt to immediate feedback with no user model; 2) *static*, that adapt based on unchanging user information; or 3) *dynamic*, that maintain user models and adapt to changing user information. To the authors' knowledge, only two robotic systems have been designed using *static* APS [12], [13]. The use of *dynamic* APS and the combination of static-dynamic adaptation have not yet been explored for persuasive robots. Dynamic adaptation has been identified as a design principle critical to the success of persuasive systems [14]. It allows for the continuous APS evolution based on changing user factors such as emotions, preferences, and skills [11]. In particular, affective state has been shown to have a substantial effect on how people process persuasive attempts [15].

In this paper, we present a novel hybrid hierarchical learning architecture that incorporates both static and dynamic user adaptation to determine effective persuasive behaviors for a socially assistive robot. We use a hierarchical system to implicitly learn individual users' 1) static cognition bias, which determines their tendency to be persuaded more by certain approaches [16], and 2) dynamic affective preferences, which influence how people process persuasive attempts [15]. Our proposed architecture incorporates two unique levels: 1) top level cognitive biases that are solved using a Contextual Bandit (CB) approach to optimize for how an individual processes information; and 2) lower level persuasive preferences optimized by *Q*-Learning (*QL*) to select a persuasive strategy for the robot to deploy that aligns with the user's affect.

The proposed architecture can be generalizable to different persuasion frameworks and applications. Herein, we consider the task of a robot persuading someone to exercise considering the Elaboration Likelihood Model (ELM) of persuasion [17]. ELM considers two different routes by which persuasion is processed: central (using rational, deliberate decisions) and peripheral (using intuitive, feeling-based decisions). To our knowledge, this research is the first to implement a static-dynamic APS for socially assistive robotics and the first to incorporate user state decomposition into persuasive HRI. Our methodology allows socially assistive robots to incorporate a more humanlike decision-making approach in order to learn to adapt to people's persuasive biases and preferences.

Manuscript received: September 9, 2021; Revised: December 3, 2021; Accepted: December 26, 2021.

This paper was recommended for publication by Editor Gentiane Venture upon evaluation of the Associate Editor and Reviewers' comments. This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Vanier Canada Graduate Scholarships (Vanier CGS), AGE-WELL, and the Canada Research Chairs (CRC) program.

Both authors are with the Autonomous Systems and Biomechatronics Lab in the Department of Mechanical and Industrial Engineering at the University of Toronto, 5 King's College Road, Toronto, ON, M5S 3G8 Canada. (Email: shane.saunderson@mail.utoronto.ca, nejat@mie.utoronto.ca).

Digital Object Identifier (DOI): see top of this page.

II. Related Work

Relevant literature reviewed here is: 1) adaptive robots [11], [18]–[23]; 2) adaptive persuasive robots [12], [13]; and 3) robot hierarchical reinforcement learning (HRL) [2], [24]–[29].

A. Adaptive Robots

There has been extensive research on (non-persuasive) adaptivity in HRI, with two recent comprehensive survey publications [11], [18]. Reactive adaptive systems have been used in social HRI applications such as matching user vocal pitch to aide in teaching [19] and mimicking nonverbal behaviors to improve social rapport [20]. Meanwhile, static adaptive systems in HRI have been used for robot-user personality matching [22] and guiding users in dressing while adapting to their mobility limitations [21].

In general, reactive HRI systems do not distinguish between individual users and instead generalize learnings across all users [11]. Meanwhile, most static adaptive HRI systems adapt based on explicitly established, a priori factors (often obtained through questionnaires) or by learning solely on interaction successes/failures without considering the influence of user context on these outcomes [11], [18].

HRI research on dynamic adaptive systems is comparatively limited [11]. Dynamic systems benefit from increased autonomy, greater robustness, and continual improvement; making them better suited to real-world, long-term interactions [18]. They have been used by robots to adapt tutoring support to a child's reading level [23] and adapt cognitive training assistance to a user's emotional arousal [2]. By leveraging both static user preferences and dynamic adaptation, robot behaviors can be adapted on deeper psychological measures, such as user cognition and emotional states, leading to a more holistic view of a user and more effective adaptation [11], [18].

B. Adaptive Persuasive Robots

To-date, robotic adaptive persuasion has been implemented as *static* APS [12], [13]. In [12], the Tangy interactive robot autonomously facilitated a Bingo game, where it used persuasive attempts to encourage user actions such as marking a Bingo card. The persuasive system selected behaviors from either a neutral, praising, suggesting, or scarcity strategy based on individual user profiles. These profiles were learned using Thompson Sampling based on the success of the different strategies in previous interactions. In [13], a Pepper robot personalized messages about healthy eating to users based on biometric assumptions of age and gender. Convolutional Neural Networks (CNNs) were used to classify users by gender and age (young, adult, or older adult). The persuasive messages were then based on the presumed goals of those age groups (e.g. younger users' goals are to improve physical appearance).

As previously mentioned both these robots use static APS, as they assume that either a user's profile [12] or age/gender [13] do not change during HRI. The Persuasive Systems Design (PSD) framework proposes design principles and evaluation methods for developing persuasive technologies [14]. Among the seven primary design principles recommended by the PSD are *personalization* – offering user-specific persuasive content – and *tailoring* – aligning persuasive content to the immediate context of a user. This framework highlights the importance of incorporating both static (personalization) and dynamic (tailoring) adaptive persuasion. To our knowledge, no robotic system has implemented dynamic APS for use in HRI nor has one explored the combination of static-dynamic adaptation.

Research in psychology has identified close to 100 different persuasive strategies ranging from altruism to deceit [30]. Within persuasive HRI, our own past research has measured the persuasiveness of strategies using objective and subjective measures including: comparing logical and emotional strategies in a robot's ability to influence participants estimates in a guessing game [31]; subjective responses on robot persuasiveness, trustworthiness, and willingness to help the robot when comparing directness and familiarity effects [32]; and comparing the effects of a robot's use of different authority types on its ability to influence user responses in attention and memory tasks [33]. Other researchers have investigated objective measures such as: robot use of goodwill, similarity, or expertise strategies in motivating users' number of exercise repetitions [34]; comparing robot use of social or factual feedback in encouraging users to use less energy on a simulated washing machine [35]; and comparing reward and coercion strategies in influencing user coffee brand selection [36].

C. Hierarchical Reinforcement Learning in Robotics

HRL has been used in social HRI to assist with cognitive training activities, such as using MAXQ for memory games [2], [24], and for activities of daily living, such as using a POMDP approach for providing reminders [25]. CB methods have been used in robot grasping applications [26] and playing bandit tasks [27], but they have not been hierarchical in nature nor used for social HRI applications.

Hybrid hierarchical control architectures using *QL* have been leveraged by mobile robots to decompose navigation problems into multiple levels with different observability. For example, in [28], a hybrid structure was used to hierarchically separate a robot navigation problem into a high level Semi-MDP with abstract task states and actions and a lower level MDP to coordinate navigation control. In [29], a hybrid structure had a high-level POMDP model of abstract, macro tasks (e.g. target search) and a lower level MDP for online navigation.

In all aforementioned robotics approaches the hierarchical structures were used to deconstruct *task spaces*. A recent survey on human-centered RL for social robotics – an approach where humans are involved in learning through either implicit or explicit rewards or guidance to the agent [37] – highlighted the potential value of using hierarchical learning architectures to deal with the complexity of *human states* and feedback [38].

Though HRL in robotics is typically used to deconstruct task spaces [2], [24]–[29], research on app-based recommender systems has shown the potential of using HRL for *user state* decomposition [39], [40]. To the authors' knowledge, no HRL approach has been used to optimize robot persuasive behaviors by deconstructing user states into hierarchical layers. Our proposed hybrid architecture is: 1) the first to incorporate dynamic states into a robotic APS in order to jointly consider static and dynamic states in a persuasive hierarchical structure, and 2) the first to consider user state decomposition in persuasive robotics applications.

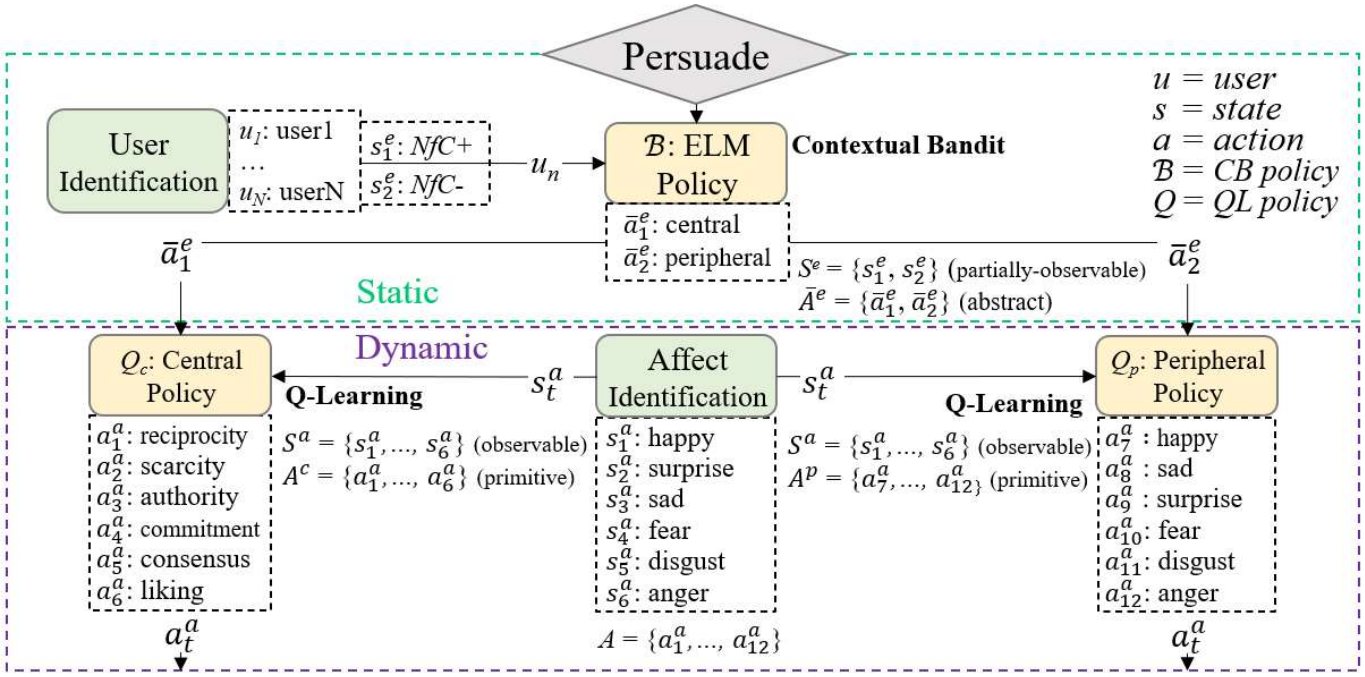

Fig. 1. HHL architecture, showing static upper layer with CB policy and dynamic lower layer with QL policies.

III. Hybrid Learning Architecture for Robot Adaptive Persuasive System

We have developed a Hybrid Hierarchical Learning (HHL) architecture for adapting persuasive behaviors of socially assistive robots during assistive HRI, Fig. 1. Our static-dynamic adaptation architecture is based upon the ELM [17], which describes how people process stimuli in order to be positively persuaded by others. The ELM considers both static, dispositional factors, and dynamic, situational factors for influencing the outcome of persuasive attempts. Within our architecture, the static dispositional state is the *Need for Cognition (*NfC*)*, a personality bias dictating the tendency for an individual to engage in and enjoy thinking [41]. Human affect is considered a dynamic, situational state [42], which can change during the course of the interaction. The ELM proposes two top level routes to persuasive processing [17]: 1) central, which focuses on rational consideration of available information, and 2) peripheral, which focuses on automatic feelings about the presented cues. Research has shown that individuals with a high NfC (*NfC+*) statistically are more biased towards the central route, which processes persuasion more logically, while those with a lower NfC (*NfC-*) scrutinize arguments less and are more biased towards the more intuitive, peripheral route [16]. With both these persuasive routes, dynamic situational factors, such as affect, can influence how people process persuasive attempts [17].

Our HHL architecture builds upon the ELM by incorporating both these static and dynamic states into a hierarchical learning system. At the top level of the hierarchy, a user's static, unobservable NfC state dictates their bias towards either the central or peripheral abstract action, Fig. 1. Though this state cannot be explicitly observed, by learning each user's persuasive preferences through rewards obtained from direct interactions, the system can implicitly learn to select the abstract action that aligns with that user's NfC bias. The lower level of the hierarchy observes user affective states in order to learn the selection of appropriate persuasive robot primitive actions. Affect is modeled using the six basic emotional states of happy, surprise, sad, fear, disgust, and anger [43]. These provide a universal framework for estimating affect from non-verbal modes like facial expressions [44] and body language [45]. In the central route, there are a possibility of 6 persuasive strategies designed based on Cialdini's principles of influence: reciprocity, commitment, consensus, liking, authority, and scarcity [46]. These principles are commonly used in persuasive technologies when logical/rational approaches are required [9], [10]. For the peripheral route, we designed 6 persuasive strategies based on the six basic emotions [43]. This emotional framework is used, herein, as the use of emotional expressions in persuasive attempts has been shown to influence how people actually process these attempts, particularly those of lower NfC [47].

We represent the robot persuasion problem using a hybrid approach, where the top layer is modeled as a CB problem and the lower level as an *QL* problem. We discuss the details of these approaches below.

A. Contextual Bandit for Static NfC Bias

We model the top level of our hierarchy as a CB problem [48], as actions (abstract or primitive) do not affect user NfC state. Each persuasive attempt made is represented by a discrete episode, *t*. Observations include the current user $u_n \in \{u_1 \dots u_N\}$ and set of available abstract actions $\bar{A}^e = \{\bar{a}_1^e, \bar{a}_2^e\}$ together with their feature vector $x_{t,a}$, referred to as the *context* [49]. The context vector provides information about both the user u_n and action $\bar{a}^e$. At each episode *t*, the system selects an abstract action $\bar{a}_t^e$ based on prior observations of rewards given the context and selected actions $(x_{1,a}, \bar{a}_1^e, r_1)$, …, $(x_{t-1,a}, \bar{a}_{t-1}^e, r_{t-1})$, and learns by observing the current reward

(provided in response to the low level primitive action taken) given the current context and selected action $(x_{t,a}, \bar{a}_t^e, r_t)$. The goal is to maximize the expected total reward $(\sum_{t=1}^{T} \boldsymbol{E}_{(x_{t,a}, r_t)})$. The policy $(\mathcal{B})$ uses Vowpal Wabbit's Contextual Bandit Reinforcement Learning [50] with an epoch-greedy exploration approach [48] to maximize r_t at each episode.

The selected abstract action $\bar{a}_t^e$ informs which lower level policy, central (Q_c) or peripheral (Q_p), will be used to determine a robot's primitive persuasive action.

B. Q-Learning for Dynamic Affective State

QL uses a Markov Decision Process (MDP) formulation to model the process of selecting a persuasive primitive action, defined by tuple $< S^a, A, R, T >$ [51]. S^a is the set of user affective states $\{s_1^a, \dots, s_6^a\}$ determined by a robot affect estimation system. A is the set of robot persuasive primitive actions, a sub-set of which is available to the central policy $A^c = \{a_1^a, \dots, a_6^a\}$ and another sub-set to the peripheral policy $A^p = \{a_7^a, \dots, a_{12}^a\}$. $R(s_t^a, a_t^a)$ is the reward function that gives a reward r_t for selecting primitive action a_t^a when observing user affective state s_t^a during episode *t*. $T(s_{t+1}^a | s_t^a, a_t^a)$ is the transition function that determines whether taking action a_t^a in state s_t^a will transition the robot into a different state s_{t+1}^a. After each episode *t*, the appropriate state-action Q-values are updated according to the Bellman equation [52]:

$$Q^*(s_t^a, a_t^a) = Q(s_t^a, a_t^a) + \alpha(r_t + \gamma \max_a Q(s_{t+1}^a, a_{t+1}^a) - Q(s_t^a, a_t^a)) \quad (1)$$

where α is the learning rate and γ is the discount factor. An epsilon-greedy exploration is utilized. r_t is determined by 3 potential outcomes, the robot: 1) persuades the user (a_g is a goal action that results in user compliance with the robot's persuasive attempt), 2) does not persuade the user and results in a user affective state transition ($s_{t+1}^a \neq s_t^a$), or 3) does not persuade user and causes no affect state transition ($s_{t+1}^a = s_t^a$):

$$R(s_t^a, a_t^a) = r_t = \begin{cases} 1, & a_g \\ -0.5, & s_{t+1}^a \neq s_t^a \\ -0.1, & s_{t+1}^a = s_t^a \end{cases} \quad (2)$$

A larger negative reward for $s_{t+1}^a \neq s_t^a$ represents a more substantial penalty for the robot's persuasive behavior failing to persuade the user and leading to a user affect change (assumed to be unwanted). This reward was kept consistent across all state changes to avoid making broad assumptions on the relative persuasive goodness/badness of specific states for each user. A smaller negative reward for $s_{t+1}^a = s_t^a$ represents a penalty for failed persuasion but no affective state change.

IV. HRI Simulated Experiments

We conducted simulated experiments to evaluate the performance of our HHL architecture in determining robot persuasive behaviors for users with different dispositional biases and situational preferences. Namely, we ran comparison studies to investigate: 1) training performance, 2) robustness with respect to longitudinal user preference change, and 3) robustness to the introduction of noisy observations.

In our HRI scenario, a socially assistive robot interacts daily with a group of 100 users for a period of one year, one individual at a time, attempting to persuade each to complete their exercise routine. We used an exercise scenario as exercise is one of the most common target applications for persuasive systems [53]. In this scenario, a robot identifies the user, estimates their affective state, and then attempts to persuade them to exercise using a selected: 1) central strategy, such as *consensus* (i.e. "*everyone else I have spoken to today has done their exercises*") or *commitment* (i.e. *"you promised me that you would complete your exercises today"*); or 2) peripheral strategy, such as happy (i.e. *"it would make me happy if you completed your daily exercises"*) or surprise (i.e. *"I'd be surprised if you didn't want to do your exercises today"*).

A. User Model

Users are first randomly assigned as either *NfC*+ or *NfC*- and only the relevant primitive actions (central for *NfC*+, peripheral for *NfC*-) are given positive, goal rewards (a_g). The distribution of goal rewards (a_g, r_t=1) is determined by the persuade/goal rate (p_g). Namely, a higher persuade rate represents a greater number of state-action pairings that lead to successful persuasion. The distribution of transition rewards ($s_{t+1}^a \neq s_t^a$, r_t=-0.5) is determined by the transition rate (p_T); a higher transition rate represents more state-action pairings that lead to user affect change. All other state-action pairings are given the default reward ($s_{t+1}^a = s_t^a$, r_t=-0.1) indicating no successful persuasion or state change.

For all transition rewards, the transition function determines a subsequent state for the user to change to ($s_{t+1}^a \neq s_t^a$). For peripheral actions leading to transition, the user's subsequent state becomes the same as the emotion the robot deployed during its previous action ($s_{t+1}^a = a_t^a$), as robot-to-person emotion contagion, the automatic transfer of affective states, has been observed in a number of social HRI studies [54]. For example, if the robot uses a persuasive strategy based on sadness that leads to a user state transition, the subsequent user affective state will be sad as well. For any central action leading to transition, we model the user's subsequent state as anger ($s_{t+1}^a = s_6^a$), since the persuasion knowledge model identifies that awareness of being persuaded increases skepticism toward the persuader and causes affective reactions of negative valence such as anger [55].

B. Comparison Method

We compare the performance of our HHL method against a flat POMDP method since a non-hierarchical formulation of the persuasion problem would consist of a partially observable state space; while affective state is observable, NfC bias is unobservable. Since the state space is discrete, we treat the POMDP as an equivalent belief MDP [56] defined by the tuple $< B, Z, A, \rho, P >$. B is the state belief space $\{b_1, \dots, b_{12}\}$ (2 NfC states x 6 affect states) informed by Z affect observations $\{z_1^a, \dots, z_6^a\} = \{s_1^a, \dots, s_6^a\}$ with available primitive actions A $=\{a_1^a, \dots, a_{12}^a\}$. At each episode *t*, the belief MDP reward function $\rho(b_t, a_t^a) = \sum_s b(s_t^a) R(s_t^a, a_t^a)$ gives the expected reward from the POMDP reward function $R(s_t^a, a_t^a)$ given over the belief state probability distribution $b(s_t^a)$. The transition function $P(z_{t+1} | b_t, a_t^a)$ determines the subsequent observation z_{t+1} for taking action a_t^a during belief state b_t. Since the state space is partially observable (e.g. the robot can observe

"happy" but not know whether a user is *NfC+* or *NfC-*), each episode must form a probabilistic belief distribution over S^a: $b(s_t^a)$. An optimal policy can be learned as a greedy policy with respect to the Bellman equation for a belief MDP [57]:

$$V_t^*(b_t) = \max_a [\sum_s b(s_t^a)\, R(s_t^a, a_t^a) + \gamma \sum_z P(z_{t+1}|b_t, a_t^a)\, V_{t-1}(b_t)] \quad (3)$$

where $V_t(b_t)$ is the value function, maximized by optimal action a_t^a, and γ is the discount factor. We use the same reward function (2) and epsilon-greedy exploration approach as HHL.

C. Experiment #1: Training Performance

To compare the learning rate of the HHL model against the POMDP, a series of 10 simulated trials were conducted with one set varying persuade rates (p_g= 10%, 20%, 30%, 40%, 50%) and one set of varying transition rates (p_T = 40%, 50%, 60%, 70%, 80%). Both methods were able to learn an optimal policy for persuading users. Learning parameters were defined to be α=0.2, γ=0.05, and ε=0.2.

1) Results

Convergence graphs comparing normalized cumulative rewards for both the HHL (blue) and POMDP (red) methods for a year of daily interactions with 100 users are shown in Fig. 2 (varying persuade rate) and Fig. 3 (varying transition rate). In general, HHL took on average 7,431 (σ=1,353) episodes to converge, versus POMDP which took on average 11,105 (σ=1,113) episodes. HHL also accumulated more average total cumulative rewards (μ=23,107; σ=2,424) than POMDP (μ=17,841; σ=1,804), indicating fewer failed persuasive episodes and faster persuasive success, regardless of persuade and transition rates. A parametric dependent t-test found this difference in total cumulative rewards to be statistically significant ($t(9)$=20.1, p<0.0001). Shapiro-Wilk tests confirmed normality of our data for both HHL ($W(10)$=0.934, p=0.49) and POMDP ($W(10)$=0.966, p=0.85).

Between trials, the total cumulative rewards obtained by the HHL increased by an average of 5.7% (σ=1.5%) per every 10% p_g increase (from 10%-50%) compared to 3.6% (σ=0.6%) for POMDP, Fig. 2. Meanwhile, HHL decreased by an average of 2.8% (σ=1.4%) per every 10% p_T increase (from 40%-80%) compared to 4.2% (σ=1.6%) for POMDP, Fig. 3.

D. Experiment #2: Robustness to User Preference Change

To test the robustness of the HHL method to changes in user preferences over time, we ran an additional 10 trials, however, in these trials we allowed user state-action preferences to change at multiple intervals (at 1 and 2 years). While NfC bias is assumed to be static over time [58], affective state preferences can change longitudinally [42]. As such, after interacting with the users for 1 year, their lower level state-action pairing preferences changed via random reassignment of reward values for each user. We randomly reassigned state-action pairing preferences again at 2 years. For example, a *NfC+* user that was initially persuaded by the consensus strategy when happy, might have changed to being persuaded by a commitment strategy at 1 year, and then changed to being persuaded by the reciprocity strategy at 2 years. However, their *NfC+* bias towards logical strategies did not change. All other parameters were the same as experiment #1.

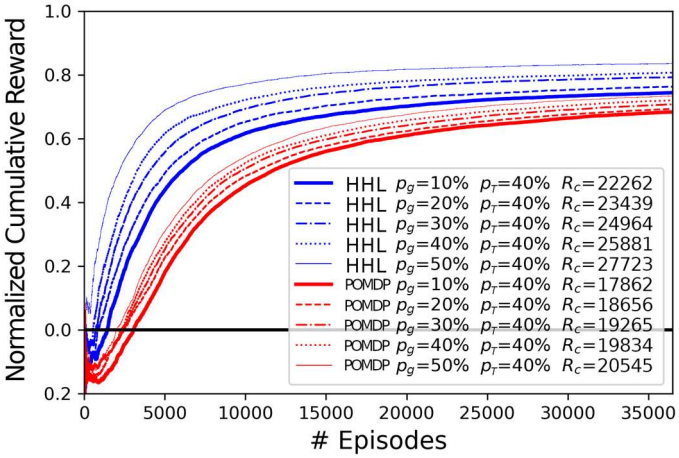

Fig. 2. Experiment #1 convergence graphs for HHL (blue) and POMDP (red) trials with varying persuade rate (p_g) and total cumulative reward (R_c).

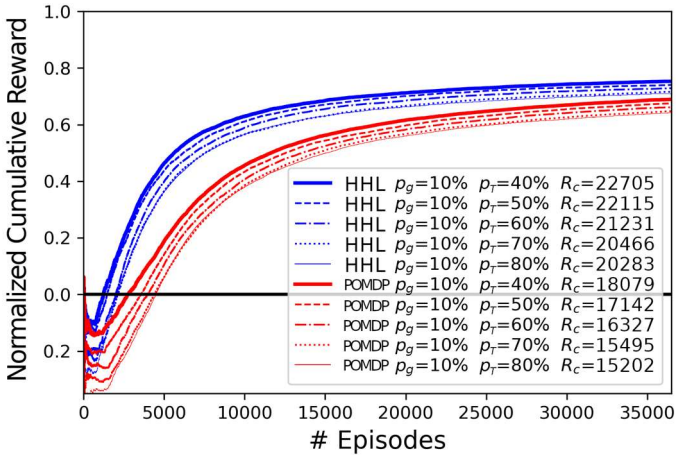

Fig. 3. Experiment #1 convergence graphs for HHL (blue) and POMDP (red) trials with varying transition rate (p_T) and total cumulative reward (R_c).

1) Results

Graphs comparing normalized cumulative rewards for both HHL (blue) and POMDP (red) methods with 100 users are shown in Fig. 4 (varying persuade rate) and Fig. 5 (varying transition rate), with preference changes visible at 1 and 2 years, respectively. We observed that after both user preference changes, our HHL method converged on average at 101,401 (σ=13,539) episodes, while POMDP converged, on average, at 130,010 (σ=18,153) episodes. HHL also accumulated more average total rewards (μ=96,315; σ=9,238) than the POMDP (μ=80,205; σ=10,136) having more successful persuasive interactions overall. A dependent t-test found this difference in total cumulative rewards to be statistically significant ($t(9)$=39.32, p<0.0001). Shapiro-Wilk tests also confirmed normality of our data for both HHL ($W(10)$=0.849, p=0.06) and POMDP ($W(10)$=0.904, p=0.24).

Between trials, the total cumulative rewards obtained by the POMDP increased by an average of 6.9% (σ=0.6%) per every 10% p_g increase (from 10%-50%) compared to 5.6% (σ=1.1%) for HHL, Fig. 4. Meanwhile, HHL decreased by an average of 1.4% (σ=1.5%) per every 10% p_T increase (from 40%-80%) compared to 3.0% (σ=1.1%) for POMDP, Fig. 5.

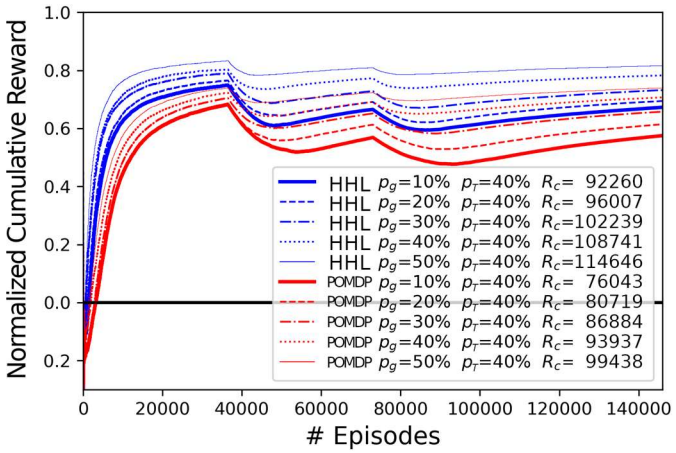

Fig. 4. Experiment #2 convergence graphs for HHL (blue) and POMDP (red) trials with varying persuade rate (p_g) and total cumulative reward (R_c).

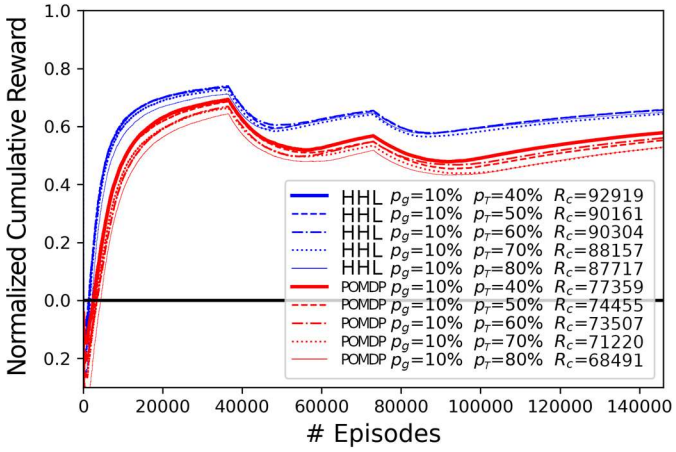

Fig. 5. Experiment #2 convergence graphs for HHL (blue) and POMDP (red) trials with varying transition rate (p_T) and total cumulative reward (R_c).

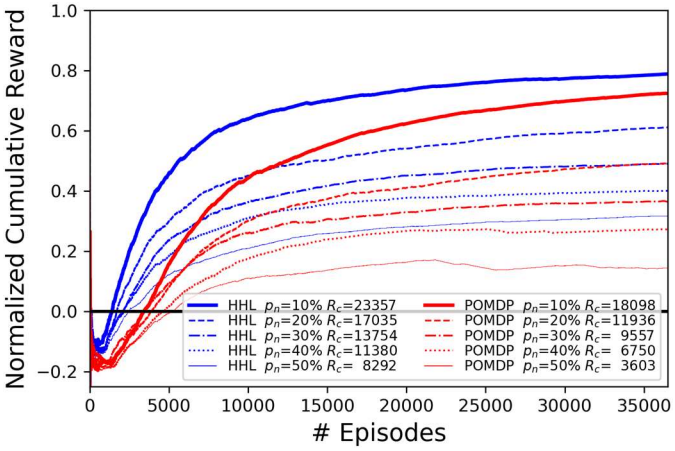

Fig. 6. Experiment #3 convergence graphs for HHL (blue) and POMDP (red) trials with varying noise rate (p_n) and total cumulative reward (R_c).

E. Experiment #3: Robustness to Noisy Observations

A third simulation was conducted to investigate the effects of noise in the affect estimation module. For these trials, the persuade and transition rates were held constant at their default values (p_g= 10%, p_T= 40%,) and affect estimation noise was introduced via a noise rate (p_n= 10%, 20%, 30%, 40%, 50%). The noise rate represented the probability of the detected user affect being randomly assigned to another affective state. All other learning parameters were consistent with experiment #1.

1) Results

A graph of normalized cumulative rewards for varying noise rates for both the HHL (blue) and POMDP (red) methods is presented in Fig. 6. In general, HHL took on average 15,283 (σ=6,549) episodes to converge, versus POMDP which took on average 22,505 (σ=9,626) episodes. HHL accumulated more cumulative rewards (μ=14,766; σ=5,773) than POMDP (μ=9,989; σ=5,500), indicating fewer failed persuasive episodes and faster persuasive convergence. A dependent t-test found the difference in total cumulative rewards to be statistically significant ($t(4)$=25.7, p<0.0001). Shapiro-Wilk tests confirmed normality of our data for both HHL ($W(5)$=0.971, p=0.88) and POMDP ($W(5)$=0.979, p=0.93).

Between trials, the total cumulative rewards obtained by the HHL decreased by an average of 22.7% (σ=5.2%) per every 10% p_n increase (from 10%-50%) compared to 32.5% (σ=11.61) for POMDP, Fig. 6.

We then conducted a test scenario where 100 simulated users were persuaded by the robot to exercise over the course of 365 days. An additional five trials were conducted with similar noise rates of 10-50%. We observed each method's persuasive success rate across the five noise rates for both the robot's first and second attempts, Table I. Our HHL method had higher average compliance rates on both the first and second attempts for all noise levels.

TABLE I COMPLIANCE RATES FOR 1ST AND 2ND PERSUASIVE ATTEMPTS

p_n	HHL 1st Attempt	HHL 2nd Attempt	POMDP 1st Attempt	POMDP 2nd Attempt
10%	93.3%	98.9%	82.4%	87.5%
20%	82.5%	91.9%	72.9%	80.9%
30%	70.8%	81.8%	52.1%	59.7%
40%	56.9%	67.0%	41.6%	49.2%
50%	47.2%	55.8%	23.9%	27.8%

V. DISCUSSIONS

Our objectives with this research were to propose a novel, hybrid hierarchical APS for HRI that could incorporate both static and dynamic user preferences, learn faster than a non-hierarchical approach, and be robust to longitudinal user changes. Our simulation results indicate a promising architecture that satisfies these objectives. Our HHL method outperformed the POMDP for all trials across all three experiments. Herein, we will discuss the detailed effects of varying persuade rate, transition rate, and noise rate on adaptive persuasion as well as some study considerations.

A. Effects of Persuasion Rate

By changing the persuade rate we simulate users who are not easily persuaded (p_g=10%) to those who are more persuadable (p_g=50%) by different robot behaviors. In experiment #1, we observed a larger average increase in cumulative reward gain for our HHL method due to increasing persuade rate compared to the POMDP method. In experiment #2, when user preferences changed, the opposite was observed: the POMDP method had a larger average increase in cumulative reward gain due to increasing persuade rate than the HHL.

The effect of increasing persuade rate on HHL had a very similar effect on rewards between the two experiments, showing that it responded more consistently to the two scenarios. The POMDP method benefitted from the increasing persuade rate for the latter experiment. This improved performance of the POMDP in experiment #2 due to increased persuade rate was presumably because at higher persuade rates, the 1 and 2 year changes were less impactful. This is because statistically, some of the preference changes may have resulted in no preference change if the new pairing randomly selected an old pairing (e.g. for p_g=10%, 1 of 6 actions led to persuasion and was statistically likely to change to one of the other 5 actions; for p_g=50%, 3 of 6 actions led to persuasion, with half a chance of being reassigned to prior pairings). Regardless, the HHL still outperformed POMDP on the least persuadable users (p_g=10%) and all other trials.

B. Effects of Transition Rate

By increasing the transition rate, we simulate users who are more likely to change their affective state due to robot persuasive actions. As transition rate increased in experiment #1, we observed a slower average decrease in cumulative rewards for the HHL model compared to the POMDP. In experiment #2, a similar but smaller decrease was observed: HHL once again has a slower average decrease in rewards compared to POMDP. This suggests that the HHL model is better at handling increased transition rates (i.e. users who more frequently change affective state due to robot actions) presumably due to the reduced state space afforded by learning a user's static NfC bias in the hierarchical structure.

C. Effects of Noise

In experiment #3, as affect noise increased, a slower average decrease in cumulative rewards was observed for our HHL method compared to the POMDP method. This suggests that the HHL model is more robust to observation noise. This is likely due to the hierarchical structure of the HHL architecture versus the non-hierarchical POMDP. For POMDP, there is only one overall policy which would directly learn from such noisy observations, whereas in the HHL, three separate policies are learned due to its hierarchy, and the noise only affects the learning of two of the policies in any given episode.

D. Considerations & Limitations

The HHL approach is a general architecture and is not limited to any one psychological framework (i.e. ELM) or set of state/action definitions. What is critical to the architecture is the hierarchical structure that separates dynamic user states in one level from static states in another to enable the use of a hybrid approach; in our case, a top level CB method and lower level *QL* method to deal with different state observabilities.

We also compared our HHL method to other ablations (e.g. considering exclusively static-only or dynamic-only user states). The HHL method performed substantially better than these ablations. The POMDP approach presented herein also performed better than the ablations and was therefore presented in this study as the most competitive benchmark for the HHL.

In our experiments, we assumed that every user state had at minimum one successful robot persuasive action associated with it. In reality, some users may be unpersuadable in certain states [59], [60]. As such, while our system tenaciously attempted to persuade a user multiple times until succeeding for learning purposes, in real-world settings, much like with human persuasion, a robot would need to learn when to give up after an appropriate number of attempts and when to try again.

VI. Conclusions

In this paper, we proposed a novel, hybrid hierarchical APS architecture for socially assistive robots engaging in persuasive HRI. The architecture uniquely separates static, dispositional user states (such as NfC) and dynamic, situational states (such as affect) in a hierarchical structure, using a CB method to optimize the top level policy for cognition bias and a *QL* method to optimize the lower level affective strategy preferences. Simulated experiments showed that the HHL converges at a faster rate than a non-hierarchical method, is more robust to longitudinal user preference change, particularly for users with a high transition rate (i.e. whose affective state can change more easily due to robot persuasive behavior), and is more robust to observation noise. Future research will incorporate our HHL methodology within the control architecture of a social robot with complementary perception and control modules to investigate the effectiveness of this proposed static-dynamic adaptive approach in real-world persuasive HRI scenarios. These modules will include user identification, affect classification, and interaction and activity monitoring, where the later will monitor the quality of the interaction and compliance/engagement of the user during adaptive persuasion.

References

[1] W. Yue, G. Louie, and G. Nejat, "A Social Robot Learning to Facilitate an Assistive Group-Based Activity from Non-expert Caregivers," *Int. J. Soc. Robot.*, vol. 12, no. 5, pp. 1159–1176, 2020.

[2] J. Chan and G. Nejat, "Social intelligence for a robot engaging people in cognitive training activities," *Int. J. Adv. Robot. Syst.*, vol. 9, pp. 1–13, 2012.

[3] R. Looije, M. A. Neerincx, and F. Cnossen, "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors," *Int. J. Hum. Comput. Stud.*, vol. 68, no. 6, pp. 386–397, 2010.

[4] M. Shao, S. F. Dos Reis, O. Ismail, X. Zhang, G. Nejat, and B. Benhabib, "You Are Doing Great ! Only One Rep Left: An Affect-Aware Social Robot for Exercising,"in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2019.

[5] E. Broadbent *et al.*, "Robots in older people's homes to improve medication adherence and quality of life: A randomised cross-over trial," *Lect. Notes Comput. Sci.*, vol. 8755, pp. 64–73, 2014.

[6] H. Nguyen and J. Masthoff, "Towards an architecture for an adaptive persuasive system," in *Proceedings of the Int. Conf. on Persuasive Technology*, 2006, pp. 108–111.

[7] M. Kaptein and A. Van Halteren, "Adaptive persuasive messaging to increase service retention: Using persuasion profiles to increase the effectiveness of email reminders," *Pers. Ubiquitous Comput.*, vol. 17, no. 6, pp. 1173–1185, 2013.

[8] S. Lukin, P. Anand, M. Walker, and S. Whittaker, "Argument strength is in the eye of the beholder: Audience effects in persuasion," in *Proceedings of the Conf. Eur. Chapter Assoc. Comput. Linguist.*, vol. 1, pp. 742–753, 2017.

[9] M. Kaptein, P. Markopoulos, B. De Ruyter, and E. Aarts, "Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles," *Int. J. Hum. Comput. Stud.*, vol. 77, pp. 38–51, 2015.

[10] M. Kaptein, B. De Ruyter, P. Markopoulos, and E. Aarts, "Adaptive persuasive systems: A study of tailored persuasive text messages to reduce snacking," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 2, pp. 1–25, 2012.

[11] G. S. Martins, L. Santos, and J. Dias, "User-Adaptive Interaction in Social Robots: A Survey Focusing on Non-physical Interaction," *Int. J.*

Soc. Robot., vol. 11, no. 1, pp. 185–205, 2019.
[12] W. Y. G. Louie and G. Nejat, "A learning from demonstration system architecture for robots learning social group recreational activities," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 808–814, 2016.
[13] B. de Carolis, F. D'Errico, and N. Macchiarulo, "'Keep the user in mind!' Persuasive effects of social robot as personalized nutritional coach," in *Proceedings of PsychoBit*, 2019, vol. 2524, pp. 1–10.
[14] H. Oinas-kukkonen and M. Harjumaa, "Persuasive Systems Design: Key Issues, Process Model, and System Features," *Commun. Assoc. Inf. Syst.*, vol. 24, no. 1, p. 28, 2009.
[15] C. R. Hullett, "The impact of mood on persuasion: A meta-analysis," *Communic. Res.*, vol. 32, no. 4, pp. 423–442, 2005.
[16] J. T. Cacioppo, R. E. Petty, F. K. Chuan, and R. Rodriguez, "Central and Peripheral Routes to Persuasion. An Individual Difference Perspective," *J. Pers. Soc. Psychol.*, vol. 51, no. 5, pp. 1032–1043, 1986.
[17] R. E. Petty and J. T. Cacioppo, "The Elaboration Likelihood Model of Persuasion," in *Communication and Persuasion*, New York, NY: Springer, 1986, pp. 123–203.
[18] M. I. Ahmad, O. Mubin, and J. Orlando, "A systematic review of adaptivity in human-robot interaction," *Multimodal Technol. Interact.*, vol. 1, no. 3, pp. 14–39, 2017.
[19] N. Lubold, E. Walker, and H. Pon-Barry, "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in *Proc. of the Int. Conf. on Human-Robot Interaction*, 2016, pp. 255–262.
[20] Q. Shen, K. Dautenhahn, J. Saunders, and H. Kose, "Can real-time, adaptive human-robot motor coordination improve humans' overall perception of a robot?," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 1, pp. 52–64, 2015.
[21] S. D. Klee, B. Q. Ferreira, R. Silva, J. P. Costeira, F. S. Melo, and M. Veloso, "Personalized assistance for dressing users," *Lect. Notes Comput. Sci.*, vol. 9388 LNCS, pp. 359–369, 2015.
[22] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *Proc. of the Int. Conf. on Human-Robot Interaction*, 2013, pp. 325–332.
[23] G. Gordon *et al.*, "Affective personalization of a social robot tutor for children's second language skills," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2016, pp. 3951–3957.
[24] J. Hemminghaus and S. Kopp, "Towards Adaptive Social Behavior Generation for Assistive Robots Using Reinforcement Learning," in *Proc. of the Int. Conf. on Human-Robot Interaction*, 2017, vol. Part F1271, pp. 332–340.
[25] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 271–281, 2003.
[26] M. Laskey *et al.*, "Multi-armed bandit models for 2D grasp planning with uncertainty," in *Proc. of the Int. Conf. on Automation Science and Engineering*, 2015, pp. 572–579.
[27] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A. Dragan, "The Assistive Multi-Armed Bandit," in *Proc. of the Int. Conf. on Human-Robot Interaction*, 2019, pp. 354–363.
[28] C. Chen, H. X. Li, and D. Dong, "Hybrid control for robot navigation - A hierarchical Q-learning algorithm," *IEEE Robot. Autom. Mag.*, vol. 15, no. 2, pp. 37–47, 2008.
[29] Y. Zhou, E. J. van Kampen, and Q. Chu, "Hybrid Hierarchical Reinforcement Learning for online guidance and navigation with partial observability," *Neurocomputing*, vol. 331, pp. 443–457, 2019.
[30] K. Kellermann and T. Cole, "Classifying Compliance Gaining Messages: Taxonomic Disorder and Strategic Confusion," *Commun. Theory*, vol. 4, no. 1, pp. 3–60, 1994.
[31] S. Saunderson and G. Nejat, "Investigating Strategies for Robot Persuasion in Social Human-Robot Interaction," *IEEE Trans. Cybern.*, pp. 1–13, 2020.
[32] S. Saunderson and G. Nejat, "Robots Asking for Favors: The Effects of Directness and Familiarity on Persuasive HRI," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1793–1800, 2021.
[33] S. Saunderson and G. Nejat, "Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction," *Sci. Robot.*, vol. 6, no. 58, pp. 1–12, 2021.
[34] K. Winkle et al, "Effective Persuasion Strategies for Socially Assistive Robots," in *Proc. of the Int. Conf. on Human-Robot Interaction*, 2019, pp. 277–285.
[35] J. Ham and C. J. H. Midden, "A Persuasive Robot to Stimulate Energy Conservation: The Influence of Positive and Negative Social Feedback and Task Similarity on Energy-Consumption Behavior," *Int. J. Soc. Robot.*, vol. 6, no. 2, pp. 163–171, 2014.
[36] M. Hashemian, M. Couto, S. Mascarenhas, A. Paiva, P. A. Santos, and R. Prada, "Persuasive social robots using reward/coercion strategies," in *Proc. of the Int. Conf. on Human-Robot Interaction*, 2020, pp. 230–232.
[37] N. Akalin and A. Loutfi, "Reinforcement learning approaches in social robotics," *Sensors*, vol. 21, no. 4, pp. 1292–1329, 2021.
[38] G. Li, R. Gomez, K. Nakamura, and B. He, "Human-Centered Reinforcement Learning: A Survey," *IEEE Trans. Human-Machine Syst.*, vol. 49, no. 4, pp. 337–349, 2019.
[39] Y. Yue, S. A. Hong, and C. Guestrin, "Hierarchical exploration for accelerating contextual bandits," in *Proc.s of the Int. Conf. on Machine Learning*, 2012, vol. 2, pp. 1895–1902.
[40] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2014, vol. 3, pp. 2296–2315.
[41] J. T. Cacioppo and R. E. Petty, "The need for cognition," *J. Pers. Soc. Psychol.*, vol. 42, no. 1, pp. 116–131, 1982.
[42] J.-B. Pavani, S. Le Vigouroux, J.-L. Kop, A. Congard, and B. Dauvier, "Affect and Affect Regulation Strategies Reciprocally Influence Each Other in Daily Life: The Case of Positive Reappraisal, Problem-Focused Coping, Appreciation, and Rumination," *J. Happiness Stud.*, vol. 17, no. 5, pp. 2077–2095, 2016.
[43] P. Ekman, "An Argument for Basic Emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.
[44] A. Mollahosseini, S. Member, B. Hasani, and S. Member, "AffectNet: A Database for Facial Expression , Valence , and Arousal Computing in the Wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2019.
[45] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI," *J. Intell. Robot. Syst. Theory Appl.*, vol. 82, no. 1, pp. 101–133, 2016.
[46] R. B. Cialdini, *Influence: Science and Practice*. Harper Collins, 1993.
[47] G. A. Van Kleef, H. van den Berg, and M. W. Heerdink, "The Persuasive Power of Emotions: Effects of Emotional Expressions on Attitude Formation and Change," *J. Appl. Psychol.*, vol. 100, no. 4, pp. 1124–1142, 2014.
[48] J. Langford and T. Zhang, "The Epoch-Greedy algorithm for contextual multi-armed bandits," in *Proc. of the Conf. on Advances in Neural Information Processing Systems*, 2009, vol. 20, no. 1, pp. 96–104.
[49] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. of the Int. Conf. on World Wide Web*, 2010, pp. 661–670.
[50] Vowpal Wabbit, "Contextual Bandit algorithms," *Github.com*, 2020. Accessed: Jun. 28, 2021. [Online]. Available: https://github.com/VowpalWabbit/vowpal_wabbit/wiki/Contextual-Bandit-algorithms.
[51] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. MIT Press, 1998.
[52] C. Watkins and P. Dayan, "Q-Learning," *Mach. Learn.*, vol. 8, no. 3–4, pp. 279–292, 1992.
[53] J. Hamari, J. Koivisto, and T. Pakkanen, "Do Persuasive Technologies Persuade? - A Review of Empirical Studies," in *Int. Conf. on Persuasive Technology*, 2014, pp. 118–136.
[54] R. Neumann and F. Strack, "'Mood contagion': The automatic transfer of mood between persons," *J. Pers. Soc. Psychol.*, vol. 79, no. 2, pp. 211–223, 2000.
[55] M. Eisend and F. Tarrahi, "Persuasion Knowledge in the Marketplace: A Meta-Analysis," *J. Consum. Psychol.*, pp. 1–20, 2021.
[56] K. P. Murphy, "A survey of POMDP solution techniques," *Environment*, vol. 2, no. September, pp. 1–32, 2000.
[57] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and Acting in Partially Observable Stochastic Domains," *Artif. Intell.*, vol. 101, no. 1–2, pp. 99–134, 1998.
[58] J. Bruinsma and R. Crutzen, "A longitudinal study on the stability of the need for cognition," *Pers. Individ. Dif.*, vol. 127, no. Jan., pp. 151–161, 2018.
[59] P. Briñol, R. E. Petty, and J. Barden, "Happiness Versus Sadness as a Determinant of Thought Confidence in Persuasion: A Self-Validation Analysis," *J. Pers. Soc. Psychol.*, vol. 93, no. 5, pp. 711–727, 2007.
[60] M. M. Mitchell, K. M. Brown, M. Morris-Villagran, and P. D. Villagran, "The effects of anger, sadness and happiness on persuasive message processing: A test of the negative state relief model," *Commun. Monogr.*, vol. 68, no. 4, pp. 347–359, 2001.