Data Science -1 (2020) 1–41
DOI 10.3233/DS-200028
IOS Press

Reinforcement learning for personalization: A systematic literature review

Floris den Hengst [a,*], Eoin Martino Grua [b,**], Ali el Hassouni [c,**] and Mark Hoogendoorn [d,**]

[a] *Dept. of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV, Amsterdam, The Netherlands*
E-mail: f.den.hengst@vu.nl; ORCID: https://orcid.org/0000-0002-2092-9904

[b] *Dept. of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV, Amsterdam, The Netherlands*
E-mail: e.m.grua@vu.nl; ORCID: https://orcid.org/0000-0002-5471-4338

[c] *Dept. of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV, Amsterdam, The Netherlands*
E-mail: a.el.hassouni@vu.nl; ORCID: https://orcid.org/0000-0003-0919-8861

[d] *Dept. of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV, Amsterdam, The Netherlands*
E-mail: m.hoogendoorn@vu.nl; ORCID: https://orcid.org/0000-0003-3356-3574

Editor: Izabela Moise (https://orcid.org/0000-0003-0370-6749)
Solicited reviews: Shihan Wang (https://orcid.org/0000-0001-5971-7522); Alessandro Rigazzi (https://orcid.org/0000-0003-2132-7726); Valerio Grossi (https://orcid.org/0000-0002-8735-5394); one anonymous reviewer

Received 2 February 2020
Accepted 3 March 2020

Abstract. The major application areas of reinforcement learning (RL) have traditionally been game playing and continuous control. In recent years, however, RL has been increasingly applied in systems that interact with humans. RL can personalize digital systems to make them more relevant to individual users. Challenges in personalization settings may be different from challenges found in traditional application areas of RL. An overview of work that uses RL for personalization, however, is lacking. In this work, we introduce a framework of personalization settings and use it in a systematic literature review. Besides setting, we review solutions and evaluation strategies. Results show that RL has been increasingly applied to personalization problems and realistic evaluations have become more prevalent. RL has become sufficiently robust to apply in contexts that involve humans and the field as a whole is growing. However, it seems not to be maturing: the ratios of studies that include a comparison or a realistic evaluation are not showing upward trends and the vast majority of algorithms are used only once. This review can be used to find related work across domains, provides insights into the state of the field and identifies opportunities for future work.

Keywords: Reinforcement learning, contextual bandits, personalization, adaptive systems, recommender systems

[*] Corresponding author. E-mail: f.den.hengst@vu.nl.
[**] Authors contributed equally.

1. Introduction

For several decades, both academia and commerce have sought to develop tailored products and services at low cost in various application domains. These reach far and wide, including medicine [5,71,77,84,178,179], human-computer interaction [61,100,118], product, news, music and video recommendations [163,164,217] and even manufacturing [39,154]. When products and services are adapted to individual tastes, they become more appealing, desirable, informative, e.g. *relevant* to the intended user than one-size-fits all alternatives. Such adaptation is referred to as *personalization* [55].

Digital systems enable personalization on a grand scale. The key enabler is data. While the software on these systems is identical for all users, the behavior of these systems can be tailored based on experiences with individual users. For example, Netflix's[1] digital video delivery mechanism includes tracking of views and ratings. These ease the gratification of diverse entertainment needs as they enable Netflix to offer instantaneous personalized content recommendations. The ability to adapt system behavior to individual tastes is becoming increasingly valuable as digital systems permeate our society.

Recently, reinforcement learning (RL) has been attracting substantial attention as an elegant paradigm for personalization based on data. For any particular environment or user state, this technique strives to determine the sequence of actions to maximize a reward. These actions are not necessarily selected to yield the highest reward *now*, but are typically selected to achieve a high reward in the long term. Returning to the Netflix example, the company may not be interested in having a user watch a single recommended video instantly, but rather aim for users to prolong their subscription after having enjoyed many recommended videos. Besides the focus on long-term goals in RL, rewards can be formulated in terms of user feedback so that no explicit definition of desired behavior is required [12,79].

RL has seen successful applications to personalization in a wide variety of domains. Some of the earliest work, such as [122,174,175] and [231] focused on web services. More recently, [107] showed that adding personalization to an existing online news recommendation engine increased click-through rates by 12.5%. Applications are not limited to web services, however. As an example from the health domain, [234] achieve optimal per-patient treatment plans to address advanced metastatic stage IIIB/IV non-small cell lung cancer in simulation. They state that 'there is significant potential of the proposed methodology for developing personalized treatment strategies in other cancers, in cystic fibrosis, and in other life-threatening diseases'. An early example of tailoring intelligent tutor behavior using RL can be found in [124]. A more recent example in this domain, [74], compared the effect of personalized and non-personalized affective feedback in language learning with a social robot for children and found that personalization significantly impacts psychological valence.

Although the aforementioned applications span various domains, they are similar in solution: they all use traits of users to achieve personalization, and all rely on implicit feedback from users. Furthermore, the use of RL in contexts that involve humans poses challenges unique to this setting. In traditional RL subfields such as game-playing and robotics, for example, simulators can be used for rapid prototyping and *in-silico* benchmarks are well established [14,21,50,97]. Contexts with humans, however, may be much harder to simulate and the deployment of autonomous agents in these contexts may come with different concerns regarding for example safety. When using RL for a personalization problem, similar issues may arise across different application domains. An overview of RL for personalization across domains, however, is lacking. We believe this is not to be attributed to fundamental differences in setting, solution or methodology, but stems from application domains working in isolation for cultural and historical reasons.

[1] https://www.netflix.com

This paper provides an overview and categorization of RL applications for personalization across a variety of application domains. It thus aids researchers and practictioners in identifying related work relevant to a specific personalization setting, promotes the understanding of how RL is used for personalization and identifies challenges across domains. We first provide a brief introduction of the RL framework and formally introduce how it can be used for personalization. We then present a framework to classify personalization settings by. The purpose of this framework is for researchers with a specific setting to identify relevant related work across domains. We then use this framework in a systematic literature review (SLR). We investigate in which settings RL is used, which solutions are common and how they are evaluated: Section 5 details the SLR protocol, results and analysis are described in Section 6. All data collected has been made available digitally [46]. Finally, we conclude with current trends challenges in Section 7.

2. Reinforcement learning for personalization

RL considers problems in the framework of *Markov decision processes* or MDPs. In this framework, an agent collects rewards over time by performing actions in an environment as depicted in Fig. 1. The goal of the agent is to maximize the total amount of collected rewards over time. In this section, we formally introduce the core concepts of MDPs and RL and include some strategies to personalization without aiming to provide an in depth introduction to RL. Following [188], we consider the related *multi-armed* and *contextual bandit* problems as special cases of the full RL problem where actions do not affect the environment and where observations of the environment are absent or present respectively. We refer the reader to [188,220] and [190] for a full introduction.

An MDP is defined as a tuple $\langle S, A, T, R, \gamma \rangle$ where $S \in \{s_1, \ldots, s_n\}$ is a finite set of states, $A \in \{a_1, \ldots, a_m\}$ a finite set of system actions, $T : S \times A \times S \to [0, 1]$ a probabilistic transition function, $R : S \times A \to \mathbb{R}$ a reward function and $\gamma \in [0, 1]$ a factor to discount future rewards. At each time step t, the system is confronted with some state s_t, performs some action a_t which yields a reward $r_{t+1} : R(s_t, a_t)$ and some state s_{t+1} following the probability distribution $T(s_t, a_t)$. A series of these states, actions and rewards from the onset to some terminal state T is called a trajectory $tr : \langle s_{t_0}, a_{t_0}, r_{t_1}, s_{t_1}, \ldots, s_{T-1}, a_{T-1}, r_T, s_T \rangle$. These trajectories typically contain the interaction histories for users with the system. A single trajectory can describe a single session of the user interacting with the system or can contain many different separate sessions. Multiple trajectories may be available in a data set $D \in \{tr_1, \ldots, tr_\ell\}$. The goal is to find a policy π^* out of all $\Pi : S \times A \to [0, 1]$ that maximizes

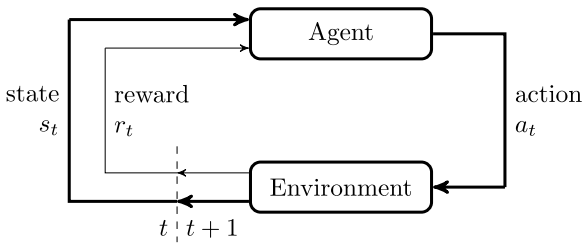

Fig. 1. The agent-environment in RL for personalization from [188].

the sum of future rewards at any t, given an end time T:

$$G_t : \sum_{k=t}^{T-1} \gamma^{k-t} r_{k+1}. \tag{1}$$

If some expectation $\mathbb{E}_\pi$ over the future reward for some policy π can be formulated, a value can be assigned to some state s given that policy:

$$V_\pi(s) = \mathbb{E}_\pi[G_t | s_t = s]. \tag{2}$$

Similarly, a value can be assigned to an action a in a state s:

$$Q_\pi(s, a) = \mathbb{E}_\pi[G_t | s_t = s, a_t = a]. \tag{3}$$

Now the optimal policy π^* should satisfy $\forall s \in S, \forall \pi \in \Pi : V_{\pi^*}(s) \geqslant V_\pi(s)$ and $\forall s \in S, a \in A, \forall \pi \in \Pi : Q_{\pi^*}(s, a) \geqslant Q_\pi(s, a)$. Assuming a suitable $\mathbb{E}_{\pi^*}[G]$, π^* consists of selecting the action that is expected to yield the highest sum of rewards:

$$\pi^*(s) = \arg\max_a Q_{\pi^*}(s, a), \quad \forall s \in S, a \in A. \tag{4}$$

With these definitions in place, we now turn to methods of finding π^*. Such methods can be categorized by considering which elements of the MDP are known. Generally, S, A and γ are determined upfront and known. T and R, on the other hand, may or may not be known. If they are both known, the expectation $\mathbb{E}_\pi[G]$ is directly available and a corresponding π^* can be found analytically. In some settings, however, T and R may be unknown and π^* must be found empirically. This can be done by estimating T, R, V, Q and finally π^* or a combination thereof using data set D. Thus, if we include approximations in Eq. (4), we get:

$$\hat{\pi^*}(s)|D = \arg\max_a \hat{Q}_{\hat{\pi^*}}(s, a)|D, \quad \forall s \in S, a \in A. \tag{5}$$

As D may lack the required trajectories for a reasonable $\mathbb{E}_{\hat{\pi^*}}[G]$ and may even be empty initially, *exploratory* actions can be selected to enrich D. Such actions need not follow $\hat{\pi^*}$ as in Eq. (5) but may be selected through some other mechanism such as sampling from the full action set A randomly.

Having introduced RL briefly, we continue by exploring some strategies in applying this framework to the problem of personalizing systems. We return to our earlier example of a video recommendation task and consider a set of n users $U \in \{u_1, \ldots, u_n\}$. A first way to adapt software systems to an individual users' needs is to define a separate environment, corresponding MDP and RL agent for each user. The overall goal becomes to find a set of optimal policies $\{\pi_1^*, \ldots, \pi_n^*\}$ for a set of environments formalized as MDPs $M : \{M_1 : \langle S_1, A_1, T_1, R_1, \gamma_1 \rangle, \ldots, M_n : \langle S_n, A_n, T_n, R_n, \gamma_n \rangle\}$. In the case of approximations as in Eq. (5), these are made per MDP based on data set D_i with trajectories only involving that environment. In the running example, videos would be recommended to a user based on previous video recommendations and selections of that particular user. The benefit of isolated MDPs is that differences between T_i and T_j or between R_i and R_j for MDPs $M_i \neq M_j$ are handled naturally, e.g. such differences do not make $\mathbb{E}_{\pi_i}[G]$ incorrect. On the other hand, similarities between T_i, T_j and R_i, R_j cannot

be used. For example, consider a video recommendation task with $S_{ij} = \{morning, afternoon, night\}$. If two users $u_i \neq u_j$ are both using a video service in the *morning* state, they may both like to watch a breakfast news broadcast whereas in the *night* state they may both prefer a talk show. Learning such patterns for each environment individually may require a substantial number of trajectories and may be infeasible in some settings, such as those where users cannot be identified across trajectories or those where each user is expected to contribute only one trajectory to D_i.

An alternative approach is to define is a single agent and MDP with user-specific information in the state space S and learn a single π^* for all users [47]. In some settings, users can be described using a function that returns a vector representation of the l features that characterize a user $\phi : U \rightarrow \langle \phi_1(U), \ldots, \phi_l(U) \rangle$. Such a vector could for example contain age, favourite genre and viewing history. If two users $u_j \neq u_i$ have both enjoyed the first "Lord of the Rings" movie and viewer u_j has followed up on a recommendation of its sequel by the system then this sequel may be a suitable recommendation for the other viewer u_i as well. Generally, this approach can be valuable when it is unclear which elements of trajectories of users u_j should be used in determining π_i^*. Conceptually, finding π^* now includes determining u_i's preference for actions given a state and determining the relationship between user preferences. This approach should therefore be able to overcome the negative transfer problem described below when enough trajectories are available. The growth in state space size, on the other hand, may require an exorbitant number of trajectories in D due to the curse of dimensionality [15]. Thus, ϕ is to be carefully designed or dimensionality reduction techniques are to be used in approaches following this strategy. As a closing remark on this approach to personalization, we note that the distinction between task-related and user-specific information is somewhat artificial as S may already contain $\phi(U)$ in many practical settings and we stress that the distinction is made for illustrative purposes here.

A third category of approaches can be considered as a middle ground between learning a single π^* and learning a π_i^* per user. It is motivated by the idea that users and corresponding environments may be similar. If this is the case, then trajectories D_j from some similar environment $M_j \neq M_i$ may prove useful in estimating $\mathbb{E}_{\pi_i}[G]$. One such an approach is based on clustering [54,75,124,191]. Formally, it requires $q \leqslant n$ groups $G \in \{g_1, \ldots, g_q\}$ and a mapping function $\Phi : M \rightarrow G$. In practice, this mapping function is typically defined on the level of users U or the feature representation $\phi(U)$. An RL agent is defined for every g_p and interacts with all environments $M_i, M_j, \Phi(M_i) = \Phi(M_j) = g_p$. Trajectories in D_i and D_j are concatenated or *pooled* to form a single D_p which is used to approximate $\mathbb{E}_{\hat{\pi_p}}[G]$ for all M_i, M_j. A combined D_p may be orders of magnitude bigger than an isolated D_i, which may result in a much better approximation $\mathbb{E}_{\hat{\pi_p}}[G]|D_p$ and a resulting $\hat{\pi_p^*}(s)|D_p$ that yields a higher reward in all environments. For example, users of the video recommendation service may be clustered by age and users in the 'infant' cluster may generally prefer children's movies over history documentaries. A related approach similarly uses trajectories D_j of other environments $M_j \neq M_i$ but still aims to find environment-specific π_i^*. Trajectories in D_j are weighted during estimation of $\mathbb{E}_{\pi_i}[G]$ using some weighting scheme. This can be understood as a generalization of the pooling approach. First, recall that $\Phi : M \rightarrow G$ for the pooling approach and note that it can be rewritten to $\Phi : M \times M \rightarrow \{0, 1\}$. The weighting scheme, now, is a generalization where $\Phi : M \times M \rightarrow \mathbb{R}$. Finding a suitable Φ can be challenging in itself and depends on the availability of user features, trajectories and the task at hand. Typical strategies are to define Φ in terms of similarity of feature representations of users $[\phi(u_i), \phi(u_j)]$ or similarity of D_i, D_j. The two previous approaches work under the assumption that T_i, T_j and R_i, R_j are similar and that Φ is suitable. If either of these assumptions is not met, pooling data may result in a policy that is suboptimal for both M_i and M_j. This phenomenon is typically referred to as the *negative transfer problem* [146].

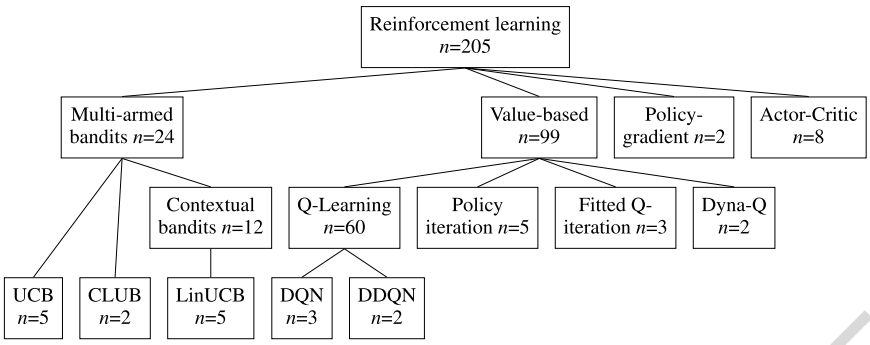

Fig. 2. Overview of types of RL algorithms discussed in this section and the number of uses in publications included in this survey. See Table 4 for a list of all (families of) algorithms used by more than one publication.

3. Algorithms

In this section we provide an overview of specific RL techniques and algorithms used for personalization. This overview is the result of our systematic literature review as can be seen in Table 4. Figure 2 contains a diagram of the discussed techniques. We start with a subset of the full RL problem known as k-armed bandits. We bridge the gap towards the full RL setting with contextual bandits approaches. Then, value-based and policy-gradient RL methods are discussed.

3.1. Multi-armed bandits

Multi-armed bandits is a simplified setting of RL. As a result, it is often used to introduce basic learning methods that can be extended to full RL algorithms [188]. In the non-associative setting, the objective is to learn how to act optimally in a single situation. Formally, this setting is equivalent to an MDP with a single state. In the associative or *contextual* version of this setting, actions are taken in more than one situation. This setting is closer to the full RL problem yet it lacks an important trait of full RL, namely that the selected action affects the situation. Both associative and non-associative multi-armed bandit approaches do not take into account temporal separation of actions and related rewards.

In general, multi-armed bandit solutions are not suitable when success is achieved by sequences of actions. Non-associative k-armed bandits solutions are only applicable when context is not important. This makes them generally unsuitable for personalizaton as it typically utilizes different personal contexts for different users by offering a different functionality. In some niche areas, however, k-armed bandits are applicable and can be very attractive due to formal guarantees on their performance. If context is of importance, contextual bandit approaches provide a good starting point for personalizing an application. These approaches hold a middle ground between non-associative multi-armed bandits and full RL solutions in terms of modeling power and ease of implementation. Their theoretical guarantees on optimality are less strong than their k-armed counterparts but they are easier to implement, evaluate and maintain than full RL solutions.

3.1.1. k-Armed bandits

In a k-armed bandit setting, one is constantly faced with the choice between k different actions [188]. Depending on the selected action, a scalar reward is obtained. This reward is drawn from a stationary probability distribution. It is assumed that an independent probability distribution exists for every action.

The goal is to maximize the expected total reward over a certain period of time. Still considering the k-armed bandit setting, we assign a value $Q(a)$ to each of the k actions and define this value as the expected reward given that the action was selected. The expected reward given that an action a is selected is defined as follows:

$$Q(a) = \mathbb{E}[r_t | a_t = a]. \tag{6}$$

In a trivial problem setting, one knows the exact value of each action and selecting the action with the highest value would constitute the optimal policy. In more realistic problems, it is fair to assume that one cannot know the values of the actions exactly. In this case, one can estimate the value of an action. We denote this estimated value with $\hat{Q}(a)$ and our goal is to have estimate $\hat{Q}(a)$ as close to the true $Q(a)$ as possible.

At each time step t, estimates of the values of actions are obtained. Always selecting the actions with the highest estimated value is called greedy action selection. In this case we are exploiting the knowledge we have built about the values of the actions. When we select actions with a lower expected value, we say we are exploring. In this case we are improving the estimates of values for these actions. In the balancing act of exploration and exploitation, we opt for exploitation to maximize the expected total reward for the next step, while opting for exploration could results in higher expected total reward in the long run.

3.1.2. Action-value methods for multi-armed bandits

Action-value methods [188] denote a collections of methods used for estimating the values of actions. The most natural way of estimating the action-values is to average the rewards that were observed. This method is called the sample-average method. The value estimate $\hat{Q}_\pi(a)$ is then defined as:

$$\hat{Q}(a) = \frac{\sum_{i=1}^{t-1} r_i \cdot \mathbb{1}_{a_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{a_i=a}}, \tag{7}$$

where $\mathbb{1}_{a_i=a}$ is 1 when $a_i = a$ is true and 0 otherwise. A default value is assigned to $\hat{Q}(a)$ when the denominator is zero. As the denominator approaches infinity, the estimate $\hat{Q}(a)$ converges to the true $Q(a)$. Again, the most basic way of selecting actions is the greedy action selection method. Here the action with the highest value is selected. In the case of a tie, one action is selected using tie-breaking methods such as random selection. Greedy action selection is defined as follows for any time point t:

$$a_t = \arg\max_a \hat{Q}(a). \tag{8}$$

Greedy action selection only exploits knowledge built up using the action-value method and only maximizes the immediate reward. This can lead to incorrect action-value approximations because actions with e.g. low *estimated* but high *actual* values are not sampled. An improvement over this greedy action selection is to randomly explore with a small probability ϵ. This method is named the ϵ-greedy action selection. A benefit of this method is that, while it is relatively simple, in the limit $\hat{Q}(a)$ will converge to $Q(a)$ [188]. This indicates that the probability of selecting the optimal action is then greater than $1 - \epsilon$ which is near certainty.

3.1.3. Incremental implementation

In Section 3.1.2 we discussed a method to estimate action-values using sample-averaging. To ensure the usability of these method in real-world applications, we need to be able to compute these values in an efficient way. Assume a setting with one action. At each iteration j a reward r_{t_j} is obtained after selecting an action. Let $\hat{Q}_n(a)$ denote the estimate value of the action after $n-1$ iterations. We can then define:

$$\hat{Q}_n(a) = \frac{r_{t_1} + r_{t_2} + r_{t_3} + \cdots + r_{t_{n-1}}}{n-1}. \tag{9}$$

Using this approach would mean storing the values of all the rewards to recalculate $\hat{Q}_n(a)$ from scratch at every iteration. There is however a more efficient way for calculating $\hat{Q}_n(a)$ that is constant in memory and computation time. Rewriting it yields the following update rule:

$$\hat{Q}_{n+1}(a) = \hat{Q}_n(a) + \frac{1}{n}\big[r_{t_n} - \hat{Q}_n(a)\big], \tag{10}$$

where the term $\hat{Q}_n(a)$ represents the old estimate, $[r_n - \hat{Q}_n(a)]$ the error in the estimate we made of the reward and $\frac{1}{n}$ the learning rate.

3.1.4. UCB: Upper-confidence bound

The greedy and ϵ-greedy action selection methods were discussed in Section 3.1.2 and it was introduced that exploration is required to establish good action-value estimates. Although ϵ-greedy explores all actions eventually, it does so randomly. A better way of exploration would take into account the action-value's proximity to the optimal value and the uncertainty in the value estimations. Intuitively, we want a selected action a to either provide a good immediate reward or else some very useful information in updating $\hat{Q}(a)$. An approach that uses this idea is the upper confidence bound action selection (UCB) method [7,65,188]. UCB is defined as follows at time step t:

$$a_t = \arg\max_a \left[\hat{Q}_n(a) + c \cdot \sqrt{\frac{\ln t}{N_t(a)}} \right], \tag{11}$$

where $N_t(a)$ is how often action a was chosen up to time t and $c > 0$ is a parameter to control the rate of exploration. The square root term denotes the level of uncertainty in the approximation of the value of action a. Hence, UCB provides an upper bound for the true value of the action a. Here, c is used to define the confidence level. When the action a is selected often, $N_t(a)$ will become larger which leads the uncertainty term to decrease. On the other hand, if the action a is not selected very often, t increases and so does the uncertainty term.

k-Armed bandit approaches address the trade-off between exploitation and exploration directly. It has been shown that the difference between the obtained rewards and optimal rewards, or the *regret*, is at best logarithmic in the number of iterations n in the absence of prior knowledge of the action value distributions and in the absence of context [102]. UCB algorithms with a regret logarithmic in and uniformly distributed over n exist [7]. This makes them a very interesting choice when strong theoretical guarantees on performance are required.

Whether these algorithms are suitable, however, depends on the setting at hand. If there is a large number of actions to choose from or when the task is not stationary k-armed bandits are typically too

simplistic. In a news recommendation task, for example, exploration may take longer than an item stays relevant. Additionally, k-armed bandits are not suitable when action values are conditioned on the situation at hand, that is: when a single action results in a different reward based on e.g. time-of-day or user-specific information such as in Section 2. In these scenarios, the problem formalization of contextual bandits and the use of function approximation are of interest.

3.1.5. Contextual bandits

In the previous sections, action-values where not associated with different situations. In this section we extend the non-associative bandit setting to the associative setting of contextual bandits. Assume a setting with n k-armed bandits problems. At each time step t one encounters a situation with a randomly selected k-armed bandits problem. We can use some of the approaches that were discussed to estimate the action values. However, this is only possible if the true action-values change slowly between the different n problems [188]. Add to this setting the fact that now at each time t a distinctive piece of information is provided about the underlying k-armed bandit which is not the actual action value. Using this information we can now learn a policy that uses the distinctive information to associate the k-armed bandit with the best action to take. This approach is called contextual bandits and uses trial-and-error to search for the optimal actions and associates these actions with situation in which they perform optimally. This type of algorithm is positioned between k-armed bandits and full RL. The similarity with RL lies in the fact that a policy is learned while the association with k-armed bandits stems from the fact that actions only affect immediate rewards. When actions are allowed to affect the next situation as well then we are dealing with RL.

3.1.6. Function approximation: LinUCB and CLUB

Despite the good theoretical characteristics of the UCB algorithm, it is not often used in the contextual setting in practice. The reason is that in practice, state and action spaces may be very large and although UCB is optimal in the uninformed case, we may do better if we use obtained information across actions and situations. Instead of maintaining isolated sample-average estimates per action or per state-action pair such as in Sections 3.1.2 and 3.1.5, we can estimate a parametric payoff function approximated from data. The parametric function takes some feature description of actions for k-armed bandit settings and state-action pairs for the contextual bandit setting and output some estimated $\hat{Q_\theta}(a)$. Here, we focus on the contextual-bandit algorithms LinUCB and CLUB.

LinUCB (Linear Upper-Confidence Bound) uses linear function approximation to calculate the confidence interval efficiently in closed form [107]. Define the expected payoff for action a with the d-dimensional featurized state $s_{t,a}$ and Θ_a^* a vector of unknown parameters as follows:

$$\mathbb{E}[r_a|s_a] = s_a^T \Theta_a^*. \tag{12}$$

Using ridge regression, an estimate of $\hat{\Theta}_a$ can be obtained [107]. Consequently, it can be shows that for any $\sigma > 0$ and $s_a \in \mathbb{R}^d$ with $\alpha = 1 + \sqrt{\ln(\frac{2}{\sigma})/2}$ a reasonably tight estimate for the expected payoff of arm a can be obtained as follows:

$$a_t = \arg\max_a \left[s_a^T \Theta_a^* + \alpha \sqrt{s_a^T A_a^{-1} s_a} \right], \tag{13}$$

where $A_a^{-1} = D_a^T D_a + I_d$ and D_a a design matrix of dimension m x d whose rows are the m contexts that are observed, $b_a \in \mathbb{R}^m$ the corresponding response vector and I_d the d x d identity matrix [107].

Similar to LinUCB, CLUB (Clustering of bandits) utilizes the linear bandit algorithm for payoff estimation [69]. In contrast to LinUCB, CLUB uses adaptive clustering in order to speed up the learning process. The main idea is to use confidence balls of user models estimate user similarity and share feedback across similar users. CLUB can thus be understood as a cluster-based alternative (see Section 2) to LinUCB algorithm.

3.2. Value-based RL

In value based RL, we learn an estimate V of the optimal value function V_{π^*} for a given policy π. We do this with the aim of finding π^*. Temporal-difference (TD) prediction is a method that learns from raw experiences without having to build a model of the environment the policy is interacting with [188,199]. In this section, we discuss various RL algorithms based on TD prediction.

3.2.1. Sarsa: On-policy temporal-difference RL

Sarsa is an on-policy temporal-difference method that learns an action-value function [181,188]. Given the current behaviour policy π, we estimate $\hat{Q}_\pi(a)$ $\forall s$, and a. This is done using transitions from state-action pair to state-action pair. Events of the form $\langle s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}\rangle$ are used in the following update rule to estimate the state-action values:

$$\hat{Q}_\pi(s_t, a_t) = \hat{Q}_\pi(s_t, a_t) + \alpha\big[r_{t+1} + \gamma\, \hat{Q}_\pi(s_{t+1}, a_{t+1}) - \hat{Q}_\pi(s_t, a_t)\big]. \tag{14}$$

This update rule is applied after every transition from s_t to s_{t+1}. In case s_{t+1} is a terminal state, a value of zero is assigned. By doing this we are ensuring that the estimate $\hat{Q}_\pi$ for a behaviour policy π while resulting in changes in π given Q_π. Sarsa will converge to an optimal action-value function Q_{π^*} and hence an optimal policy π^* in the limit given that all possible state-action pairs are visited an infinite amount of time [188]. Consequently, Sarsa converges to the greedy policy in the limit. Algorithm 1 shows Sarsa in more detail.

3.2.2. Q-learning: Off-policy temporal-difference RL

Q-learning was one of the breakthroughs in the field of RL [188,219]. Q-learning is classified as an off-policy temporal-difference algorithm for control. Similar to Sarsa, Q-learning approximates the optimal

Algorithm 1: Sarsa – An on-policy temporal-difference RL algorithm

Parameters: learning rate $\alpha \in (0, 1]$ and $\epsilon > 0.0$;
Initialize $\hat{Q}_\pi$ $\forall\, s \in S, a \in A$. For terminal states initialize the value with 0. **for** *each episode* **do**
  Initialize s
  Choose action a in s using π derived from $\hat{Q}_\pi$ (e.g. ϵ-*greedy*)
  for *each step in episode* **do**
    Select action a and obtain reward r and next state s'
    Take next action a' from s' following π derived from $\hat{Q}_\pi$ (e.g. ϵ-*greedy*)
    $\hat{Q}_\pi(s, a) = \hat{Q}_\pi(s, a) + \alpha[r + \gamma\, \hat{Q}_\pi(s', a') - \hat{Q}_\pi(s, a)]$
    Set $s = s'$ and $a = a'$
    Stop loop if s is terminal
  end
end

Algorithm 2: Q-Learning – An off-policy temporal-difference RL algorithm

Parameters: learning rate $\alpha \in (0, 1]$ and $\epsilon > 0$.
Initialize $\hat{Q}_\pi \; \forall \; s \in S, a \in A$. For terminal states initialize the value with 0. **for** *each episode* **do**
  Initialize s

  for *each step in episode* **do**
    Choose action a in s using π derived from $\hat{Q}_\pi$ (e.g. ϵ-*greedy*)
    Take action a and obtain reward r and next state s'
    $\hat{Q}_\pi(s, a) = \hat{Q}_\pi(s, a) + \alpha[r + \gamma \cdot \arg\max_a \hat{Q}_\pi(s', a) - \hat{Q}_\pi(s, a)]$
    Set $s = s'$
    Stop loop if s is terminal
  end
end

action-value function Q_{π^*} by learning $\hat{Q_{\pi^*}}$. Differently from Sarsa, Q-learning learns $\hat{Q_{\pi^*}}$ independently of the policy being followed. The policy being followed still has an effect on the learning process, but only by determining which state-action pairs are visited and consequently updated. Algorithm 2 shows Q-learning in more detail. The update rule for Q-learning is defined as follows:

$$\hat{Q}_\pi(s_t, a_t) = \hat{Q}_\pi(s_t, a_t) + \alpha\big[r_{t+1} + \gamma \max_a \hat{Q}_\pi(s_{t+1}, a) - \hat{Q}_\pi(s_t, a_t)\big]. \tag{15}$$

3.2.3. *Value-function approximation*

In Sections 3.2.2 and 3.2.1 we discussed tabular algorithms for value-based RL. In this section we discuss function approximation in RL for estimating state-value functions from a known policy π (i.e. on-policy RL). The difference with the tabular approach is that we represent v_π as a parameterized function with a weight vector $w \in \mathbb{R}^d$ where $\hat{v}(s, w) \approx v_\pi(s)$ is the approximated value of state s given the learned weights w. Different function approximators can be used to estimate $\hat{v}$. For instance, $\hat{v}$ can be a deep neural network with w representing the weights of the network. In the tabular version of value-based RL, states and their estimated values are isolated from each other while in function approximation adjusting one weight in the network can lead to changes in the estimated values of many states. This form of learning is powerful due its ability to generalize across different states, but at the same time may lead to more complex models that are harder to understand and to tune. An example of value-function approximation is the deep Q-network (DQN) algorithm [133]. This algorithm combines deep (convolutional) neural network and Q-learning. Using DQN, it was shown that RL agents can achieve state-of-the-art performances on many problems without relying on engineered features. DNQ learns directly from raw (pixel) data instead. The following update rule is an alteration of the Q-learning (semi-gradient of Q-learning [188]) update rule for estimating the weights of the network:

$$w_{t+1} = w_t + \alpha\Big[r_{t+1} + \gamma \cdot \max_a \hat{Q}_\pi(s_{t+1}, a, w_t) - \hat{Q}_\pi(s_t, a_t, w_t)\Big]\nabla_{wt}\hat{Q}_\pi(s_t, a_t, w_t). \tag{16}$$

3.3. *Policy-gradient RL*

In value-based RL values of actions are approximated and then a policy is derived by selecting actions using a certain selection strategy. In policy-gradient RL we learn a parameterized policy directly [188,

189]. Consequently, we can select actions without the need for an explicit value function. Let $\Theta \in \mathbb{R}^d$ where d is the dimension of the parameter vector Θ. For policy-based methods that also rely on a value function, we denote the function's weight vector denoted by $w \in \mathbb{R}^{d'}$ as $\hat{v}(s, w)$. Define the probability of selecting action a at time step t given state s with policy parameters Θ as:

$$\pi(a|s, \Theta) = P[a_t = a|s_t = s, \Theta_t = \Theta]. \tag{17}$$

Consider a function $J(\Theta)$ that quantifies the performance of the policy π with respect to parameter vector Θ. The goal is to optimize Θ such that $J(\Theta)$ is maximized. We use the following update rule to approximate gradient ascent in J where the term $\widehat{\nabla J(\Theta_t)} \in \mathbb{R}^d$ approximates the gradient of $J(\Theta)$ at t:

$$\Theta_{t+1} = \Theta_t + \alpha \widehat{\nabla J(\Theta_t)}. \tag{18}$$

3.4. Actor-critic

In actor-critic methods [98,188] both the value and policy functions are approximated. The actor in actor-critic is the learned policy while the critic approximates the value function. Algorithm 3 shows the one-step episodic actor-critic algorithm in more detail. The update rule for the parameter vector Θ is defined as follows:

$$\Theta_{t+1} = \Theta_t + \alpha \delta_t \frac{\nabla \pi(a|s_t, \Theta_t)}{\pi(a|s_t, \Theta_t)}, \tag{19}$$

where δ_t is defined as follows:

$$\delta_t = r_{t+1} + \gamma \hat{v}(s_{t+1}, w) - \hat{v}(s_t, w). \tag{20}$$

Algorithm 3: One-step episodic actor-critic

Input: a differentiable policy $\pi(a|s, \Theta)$
Input: a differentiable state-value function $\hat{v}(s, w)$
Parameters: $\alpha(\Theta) > 0$ and $\alpha(w) > 0$ Initialize $\Theta \in \mathbb{R}^d$ and $w \in \mathbb{R}^{d'}$
for *each episode* **do**
  Initialize S
  I = 1
  for *each step in episode* **do**
    Choose action a in s using π: $a \sim \pi(.|s, \Theta)$
    Take action a and obtain reward r and next state s'
    $\delta = r + \gamma \hat{v}(s', w) - \hat{v}(s, w)$
    $w = w + \alpha(w)\delta \nabla \hat{v}(s, w)$
    $\Theta = \Theta + \alpha(\Theta) I \delta \nabla \ln \pi(a|s, \Theta)$
    $I = \gamma I$
    $s = s'$
  end
end

Table 1
Framework to categorize personalization setting by

Category	A#	Aspect	Description	Range
Suitability outcome	A1	Control	The extent to which the user defines the suitability of behavior explicitly.	Explicit – implicit
	A2	Safety	The extent to which safety is of importance.	Trivial – critical
Upfront knowledge	A3	User models	The a priori availability of models that describe user responses to system behavior.	Unavailable – unlimited
	A4	Data availability	The a priori availability of human responses to system behavior.	Unavailable – unlimited
New Experiences	A5	Interaction availability	The availability of new samples of interactions with individuals.	Unavailable – unlimited
	A6	Privacy sensitivity	The degree to which privacy is a concern.	Trivial – critical
	A7	State observability	The degree to which all information to base personalization can be measured.	Partial – full

4. A classification of personalization settings

Personalization has many different definitions [30,55,165]. We adopt the definition proposed in [55] as it is based on 21 existing definitions found in literature and suits a variety of application domains: "personalization is a process that changes the functionality, interface, information access and content, or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals". This definition identifies personalization as a process and mentions an existing system subject to that process. We include aspects of both the desired process of change and existing system in our framework. Section 5.4 further details how this framework was used in a SLR.

Table 1 provides an overview of the framework. On a high level, we distinguish three categories. The first category contains aspects of suitability of system behavior. We differentiate settings in which suitability of system behavior is determined explicitly by users and settings in which it is inferred by the system after observing user behavior [172]. For example, a user can explicitly rate suitability of a video recommendation; a system can also infer suitability by observing whether the user decides to watch the video. Whether implicit or explicit feedback is preferable depends on availability and quality of feedback signals [89,143,172]. Besides suitability, we consider safety of system behavior. Unaltered RL algorithms use trial-and-error style exploration to optimize their behavior yet this may not suit a particular domain [78,92,136,153]. For example, tailoring the insulin delivery policy of an artificial pancreas to the metabolism of an individual requires trial insulin delivery action but these should only be sampled when their outcome is within safe certainty bounds [44]. If safety is a significant concern in the systems' application domain, specifically designed safety-aware RL techniques may be required, see [149] and [64] for overviews of such techniques.

Aspects in the second category deal with the availability of upfront knowledge. Firstly, knowledge of how users respond to system actions may be captured in user models. Such models open up a range of RL solutions that require less or no sampling of new interactions with users [81]. As an example, user pain models are used to predict suitability of exercises in an adaptive physical rehabilitation curriculum manager a priori [208]. Models can also be used to interact with the RL agent in simulation. For example, dialogue agent modules may be trained by interacting with a simulated chatbot user [47,95,105]. Secondly, upfront knowledge may be available in the form of data on human responses to system behavior. This data can be used to derive user models and can be used to optimize policies directly and provide high-confidence evaluations of such policies [111,202–204].

The third category details new experiences. Empirical RL approaches have proven capable of modelling extremely complex dynamics, however, this typically requires complex estimators that in turn need substantial amounts of training data. The availability of users to interact with is therefore a major consideration when designing an RL solution. A second aspect that relates to the use of new experiences is privacy sensitivity of the setting. Privacy sensitivity is of importance as it may restrict sharing, pooling or any other specific usage of data [9,76]. Finally, we identify the state observability as a relevant aspect. In some settings, the true environment state cannot be observed directly but must be estimated using available observations. This may be common as personalization exploits differences in mental [22,96,217] and physical state [67,125]. For example, recommending appropriate music during running involves matching songs to the user emotional state and e.g. running pace. Both mental and physical state may be hard to measure accurately [2,17,152].

Although aspects in Table 1 are presented separately, we explicitly note that they are not mutually independent. Settings where privacy is a major concern, for example, are expected to typically have less existing and new interactions available. Similarly, safety requirements will impact new interaction availability. Presence of upfront knowledge is mostly of interest in settings where control lies with the system as it may ease the control task. In contrast, user models may be marginally important if desired behavior is specified by the user in full. Finally, a lack of upfront knowledge and partial observability complicates adhering to safety requirements.

5. A systematic literature review

A SLR is 'a form of secondary study that uses a well-defined methodology to identify, analyze and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable' [23]. PRISMA is a standard for reporting on SLRs and details eligibility criteria, article collection, screening process, data extraction and data synthesis [135]. This section contains a report on this SLR according to the PRISMA statement. This SLR was a collaborative work to which all authors contributed. We denote authors by abbreviation of their names, e.g. FDH, EG, AEH and MH.

5.1. Inclusion criteria

Studies in this SLR were included on the basis of three eligibility criteria. To be included, articles had to be published in a peer-reviewed journal or conference proceedings in English. Secondly, the study had to address a problem fitting to our definition of personalization as described in Section 4. Finally, the study had to use a RL algorithm to address such a personalization problem. Here, we view contextual bandit algorithms as a subset of RL algorithms and thus included them in our analysis. Additionally, we excluded studies in which a RL algorithm was used for purposes other than personalization.

5.2. Search strategy

Figure 3 contains an overview of the SLR process. The first step is to run a query on a set of databases. For this SLR, a query was run on Scopus, IEEE Xplore, ACM's full-text collection, DBLP and Google Scholar on June 6, 2018. These databases were selected as their combined index spans a wide range, and their combined result set was sufficiently large for this study. Scopus and IEEE Xplore support queries on title, keywords and abstract. ACM's full-text collection, DBLP and Google scholar do not support queries on keywords and abstract content. We therefore ran two kinds of queries: we queried

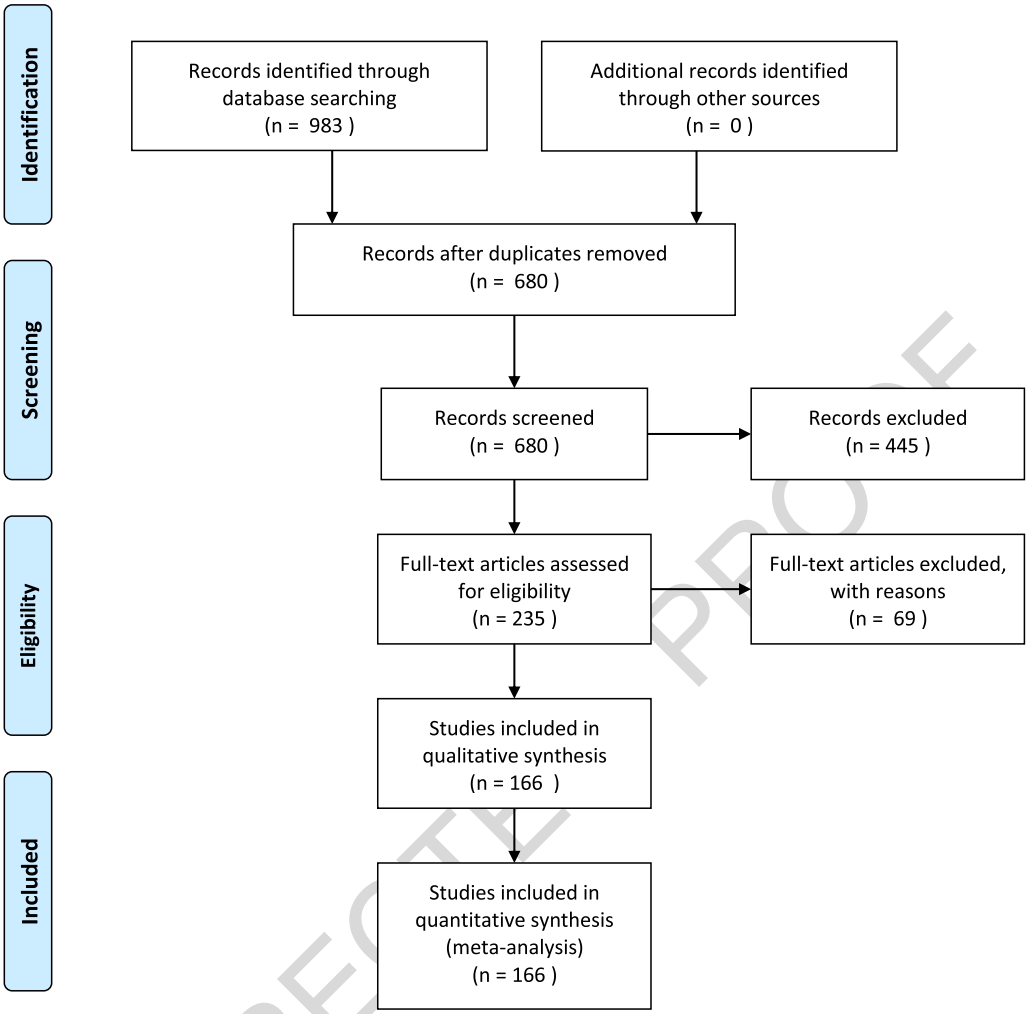

Fig. 3. Overview of the SLR process.

on title only for ACM's full-text collection, DBLP and Google Scholar and we extended this query to keywords and abstract content for Scopus and IEEE Xplore. The query was constructed by combining techniques of interest and keywords for the personalization problem. For techniques of interest the terms 'reinforcement learning' and 'contextual bandits' were used. For the personalization problem, variations on the words 'personalized', 'customized', 'individualized' and 'tailored' were included in British and American spelling. All queries are listed in Appendix A. Query results were de-duplicated and stored in a spreadsheet.

5.3. Screening process

In the screening process, all query results are tested against the inclusion criteria from Section 5.1 in two phases. We used all criteria in both phases. In the first phase, we assessed eligibility based on keywords, abstract and title whereas we used full text of the article in the second phase. In the first phase, a spreadsheet with de-duplicated results was shared with all authors via Google Drive. Studies

were assigned randomly to authors who scored each study by the eligibility criteria. The results of this screening were verified by one of the other authors, assigned randomly. Disagreements were settled in meetings involving those in disagreement and FDH if necessary. In addition to eligibility results, author preferences for full-text screening were recorded on a three-point scale. Studies that were not considered eligible were not taken into account beyond this point, all other studies were included in the second phase.

In the second phase, data on eligible studies was copied to a new spreadsheet. This sheet was again shared via Google Drive. Full texts were retrieved and evenly divided amongst authors according to preference. For each study, the assigned author then assessed eligibility based on full text and extracted the data items detailed below.

5.4. Data items

Data on setting, solution and methodology were collected. Table 2 contains all data items for this SLR. For data on setting, we operationalized our framework from Table 1 in Section 4. To assess trends in solution, algorithms used, number of MDP models (see Section 2) and training regime were recorded. Specifically, we noted whether training was performed by interacting with actual users ('live'), using existing data and a simulator of user behavior. For the algorithms, we recorded the name as used by the authors. To gauge maturity of the proposed solutions and the field as a whole, data on the evaluation strategy and baselines used were extracted. Again, we listed whether evaluation included 'live' interaction with users, existing interactions between systems and users or using a simulator. Finally, publication year and application domain were registered to enable identification of trends over time and across domains. The list of domains was composed as follows: during phase one of the screening process, all authors recorded a domain for each included paper, yielding a highly inconsistent initial set of domains. This set was simplified into a more consistent set of domains which was used during full-text screening. For papers that did not fall into this consistent set of domains, two categories were added: a 'Domain Independent' and an 'Other' category. The actual domain was recorded for the five papers in the 'Other' category. These domains were not further consolidated as all five papers were assigned to unique domains not encountered before.

5.5. Synthesis and analysis

To facilitate analysis, reported algorithms were normalized using simple text normalization and key-collision methods. The resulting mappings are available in the dataset release [46]. Data was summarized using descriptive statistics and figures with an accompanying narrative to gain insight into trends with respect to settings, solutions and evaluation over time and across domains.

6. Results

The quantitative synthesis and analyses introduced in Section 5.5 were applied to the collected data. In this section, we present insights obtained. We focus on the major insights and encourage the reader to explore the tabular view in Appendix B or the collected data for further analysis [46].

Before diving into the details of the study in light of the classification scheme we have proposed, let us first study some general trends. Figure 4 shows the number of publications addressing personalization using RL techniques over time. A clear increase can be seen. With over forty entries, the health domain

Table 2
Data items in SLR. The last column relates data items to aspects of setting from Table 1 where applicable

Category	#	Data item	Values	A#
Setting	1	User defines suitability of system behavior explicitly	Yes, No	A1
	2	Suitability of system behavior is derived	Yes, No	A1
	3	Safety is mentioned as a concern in the article	Yes, No	A2
	4	Privacy is mentioned as a concern in the article	Yes, No	A6
	5	Models of user responses to system behavior are available	Yes, No	A3
	6	Data on user responses to system behavior are available	Yes. No	A4
	7	New interactions with users can be sampled with ease	Yes, No	A5
	8	All information to base personalization on can be measured	Yes, No	A7
Solution	9	Algorithms	N/A	–
	10	Number of learners	1, 1/user, 1/group, multiple	–
	11	Usage of traits of the user	state, other, not used	–
	12	Training mode	online, batch, other, unknown	–
	13	Training in simulation	Yes, No	A3
	14	Training on a real-life dataset	Yes, No	A4
	15	Training in 'live' setting	Yes, No	A5
Evaluation	16	Evaluation in simulation	Yes, No	A3
	17	Evaluation on a real-life dataset	Yes, No	A4
	18	Evaluation in 'live' setting	Yes, No	A5
	19	Comparison with 'no personalization'	Yes, No	–
	20	Comparison with non-RL methods	Yes, No	–

contains by far the most articles, followed by entertainment, education and commerce with all approximately just over twenty five entries. Other domains contain less than twelve papers in total. Figure 5(a) shows the popularity of domains for the five most recent years and seems to indicate that the number of articles in the health domain is steadily growing, in contrast with the other domains. Of course, these graphs are based on a limited number of publications, so drawing strong conclusions from these results is difficult. We do need to take into account that the popularity of RL for personalization is increasing in general. Therefore Fig. 5(b) shows the relative distribution of studies over domains for the five most recent years. Now we see that the health domain is just following the overall trend, and is not becoming more popular within studies that use RL for personalization. We fail to identify clear trends for other domains from these figures.

6.1. Setting

Table 3 provides an overview of the data related to setting in which the studies were conducted. The table shows that user responses to system behavior are present in a minority of cases (66/166). Additionally, models of user behavior are only used in around one quarter of all publications. The suitability of system behavior is much more frequently derived from data (130/166) rather than explicitly collected by users (39/166). Privacy is clearly not within the scope of most articles, only in 9 out of 166 cases do we see this issue explicitly mentioned. Safety concerns, however, are mentioned in a reasonable proportion of studies (30/166). Interactions can generally be sampled with ease and the resulting information is frequently sufficient to base personalization of the system at hand on.

Let us dive into some aspects in a bit more detail. A first trend we anticipate is an increase of the fraction of studies working with real data on human responses over the years, considering the digitization

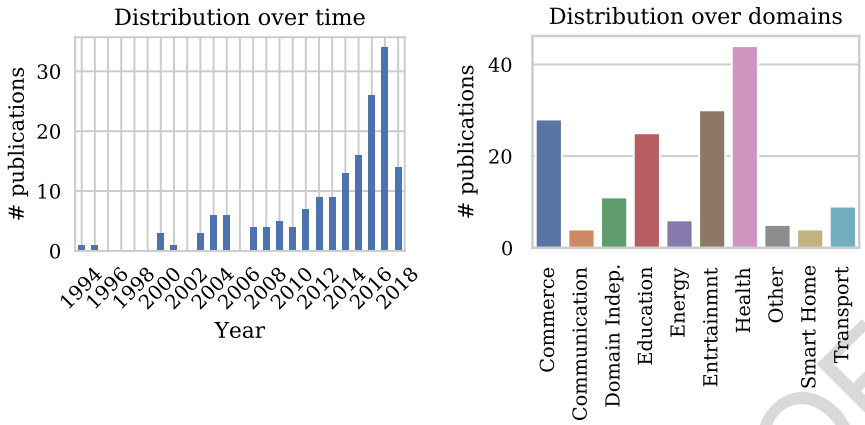

Fig. 4. Distribution of included papers over time and over domains. Note that only studies published prior to the query date of June 6, 2018 were included.

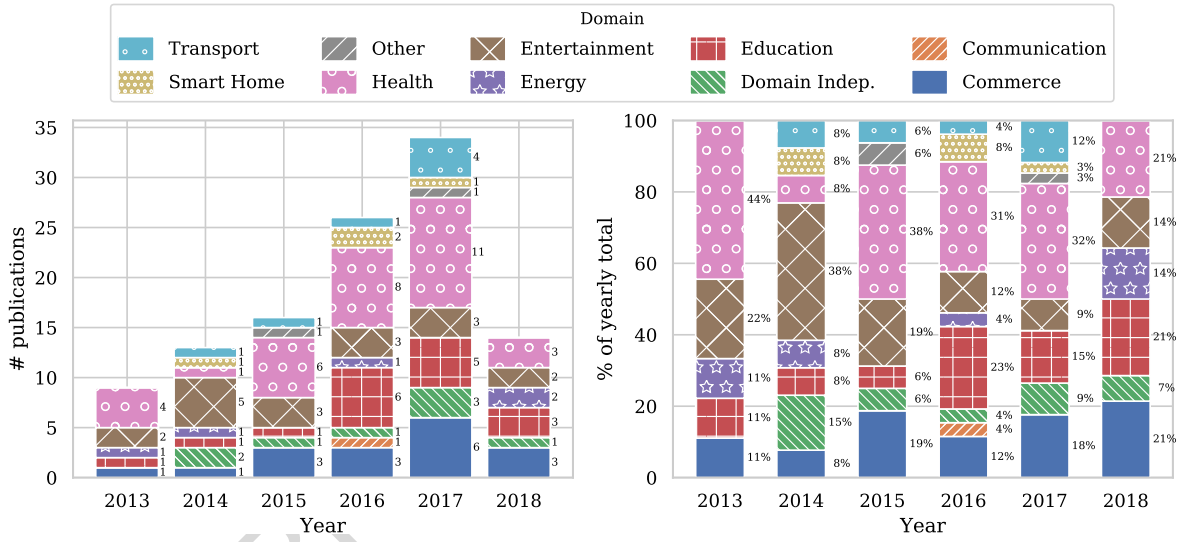

Fig. 5. Popularity of domains for the five most recent years.

Table 3
Number of publications by aspects of setting

Aspect	#
User defines suitability of system behavior explicitly	39
Suitability of system behavior is derived	130
Safety is mentioned as a concern in the article	30
Privacy is mentioned as a concern in the article	9
Models of user responses to system behavior are available	41
Data on user responses to system behavior are available	66
New interactions with users can be sampled with ease	97
All information to base personalization on can be measured	132

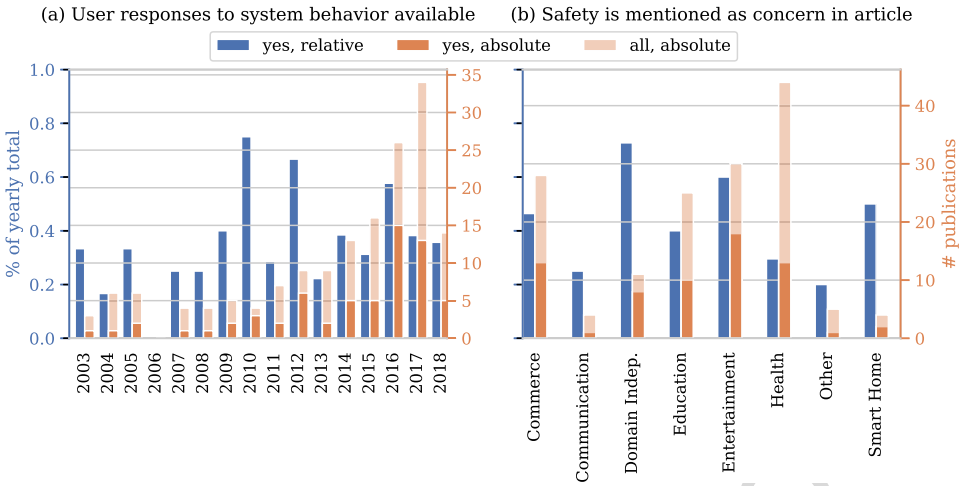

Fig. 6. Availability of user responses over time (a), and mentions of safety as a concern over domains (b).

trend and associated data collection. Figure 6(a) shows the fraction of papers for which data on user responses to system behavior is available over time. Surprisingly, we see that this fraction does not show any clear trend over time. Another aspect of interest relates to safety issues in particular domains. We hypothesize that in certain domains, such as health, safety is more frequently mentioned as a concern. Figure 6(b) shows the fraction of papers of the different domains in which safety is mentioned. Indeed, we clearly see that certain domains mention safety much more frequently than other domains. Third, we explore the ease with which interactions with users can be sampled. Again, we expect to see substantial differences between domains. Figure 7 confirms our intuition. Interactions can be sampled with ease more frequently in studies in the commerce, entertainment, energy, and smart homes domains when compared to communication and health domains.

Finally, we investigate whether upfront knowledge is available. In our analysis, we explore both real data as well user models being available upfront. One would expect papers to have at least one of these two prior to starting experiments. User models and not real data were reported in 41 studies, while 53 articles used real data but no user model and 12 use both. We see that for 71 studies neither is available. In roughly half of these, simulators were used for both training (38/71) and evaluation (37/71). In a minority, training (15/71) and evaluation (17/71) were performed in a live setting, e.g. while collecting data.

6.2. *Solution*

In our investigation into solutions, we first explore the algorithms that were used. Figure 8 shows the distribution of usage frequency. A vast majority of the algorithms are used only once, some techniques are used a couple of times and one algorithm is used 60 times. Note again that we use the name of the algorithms used by the authors as a basis for this analysis. Table 4 lists the algorithms that were used more than once. A significant number of studies (60/166) use the Q-learning algorithm. At the same time, a substantial number of articles (18/166) reports the use of RL as the underlying algorithmic framework without specifying an actual algorithm. The contextual bandits, Sarsa, actor-critic and inverse RL (IRL) algorithms are used in respectively (18/166), (12/166), (8/166), (8/166) and (7/166) papers. We also observe some additional algorithms from the contextual bandits family, such as UCB

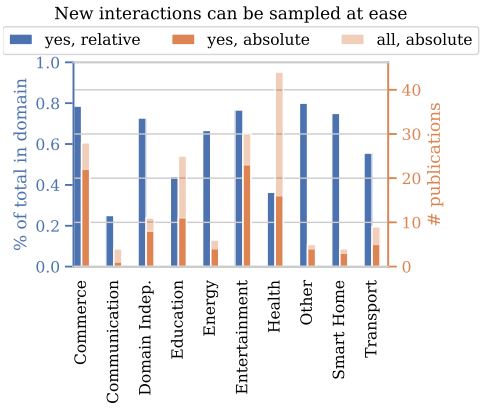

Fig. 7. New interactions with users can be sampled with ease.

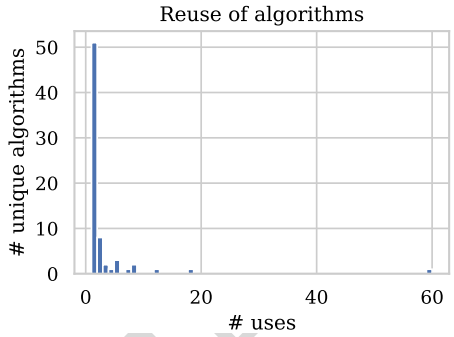

Fig. 8. Distribution of algorithm usage frequencies.

and LinUCB. Furthermore, we find various mentions that indicate the usage of deep neural networks: deep reinforcement learning, DQN and DDQN. In general, we find that some publications refer to a specific algorithm whereas others only report generic techniques or families thereof.

Figure 9(a) lists the number of models used in the included publications. The majority of solutions relies on a single-model architecture. On the other end of the spectrum lies the architecture of using one model per person. This architecture comes second in usage frequency. The architecture that uses one model per group can be considered a middle ground between these former two. In this architecture, only experiences with relevant individuals can be shared. Comparisons between architectures are rare. We continue by investigating whether and where traits of the individual were used in relation to these architectures. Table 5 provides an overview. Out of all papers that use one model, 52.7% did not use the traits of the individuals and 41.7 % included traits in the state space. 47.5% of the papers include the traits of the individuals in the state representation while in 37.3% of the papers the traits were not included. In 15.3% of the cases this was not known.

Figure 9(b) shows the popularity of using a simulator for training per domain. We see that a substantial percentage of publications use a simulator and that simulators are used in all domains. Simulators are used in the majority of publications for the energy, transport, communication and entertainment domains. In publications in the first three out of these domains, we typically find applications that require large-scale implementation and have a big impact on infrastructure, e.g. control of the entire energy grid or

Table 4
Algorithm usage for all algorithms that were used in more than one publication

Algorithm	# of uses
Q-learning [219]	60
RL, not further specified	18
Contextual bandits	12
Sarsa [187]	8
Actor-critic	8
Inverse reinforcement learning	7
UCB [7]	5
Policy iteration	5
LinUCB [37]	5
Deep reinforcement learning	4
Fitted Q-iteration [166]	3
DQN [133]	3
Interactive reinforcement learning	2
TD-learning	2
DYNA-Q [186]	2
Policy gradient	2
CLUB [69]	2
Monte Carlo	2
Thompson sampling	2
DDQN [212]	2

a fleet of taxis in a large city. This complicates the collection of useful realistic dataset and training in a live setting. This is not the case for the entertainment domain with 17 works using a simulator for training. Further investigation shows that nine out of these 17 also include training on real data or in a 'live' setting. It seems that training on a simulator is part of the validation of the algorithm rather than the prime contribution of the paper in the entertainment domain.

6.3. Evaluation

In investigating evaluation rigor, we first turn to the data on which evaluations are based. Figure 10 shows how many studies include an evaluation in a 'live' setting or using existing interactions with users. In the years up to 2007 few studies were done and most of these included realistic evaluations. In more recent years, the absolute number of studies shows a marked upward trend to which the relative number of articles that include a realistic evaluation fails to keep pace. Figure 10 also shows the number of realistic evaluations per domain. Disregarding the smart home domain, as it contains only four studies, the highest ratio of real evaluations can be found in the commerce and entertainment domains, followed by the health domain.

We look at possible reasons for a lack of realistic evaluation using our categorization of settings from Section 4. Indeed, there are 63 studies with no realistic evaluation versus 104 with a realistic evaluation. Because these group sizes differ, we include ratios with respect to these totals in Table 6. The biggest difference between ratios of studies with and without a realistic evaluation is in the upfront availability of data on interactions with users. This is not surprising, as it is natural to use existing interactions for evaluation when they are available already. The second biggest difference between the groups is whether

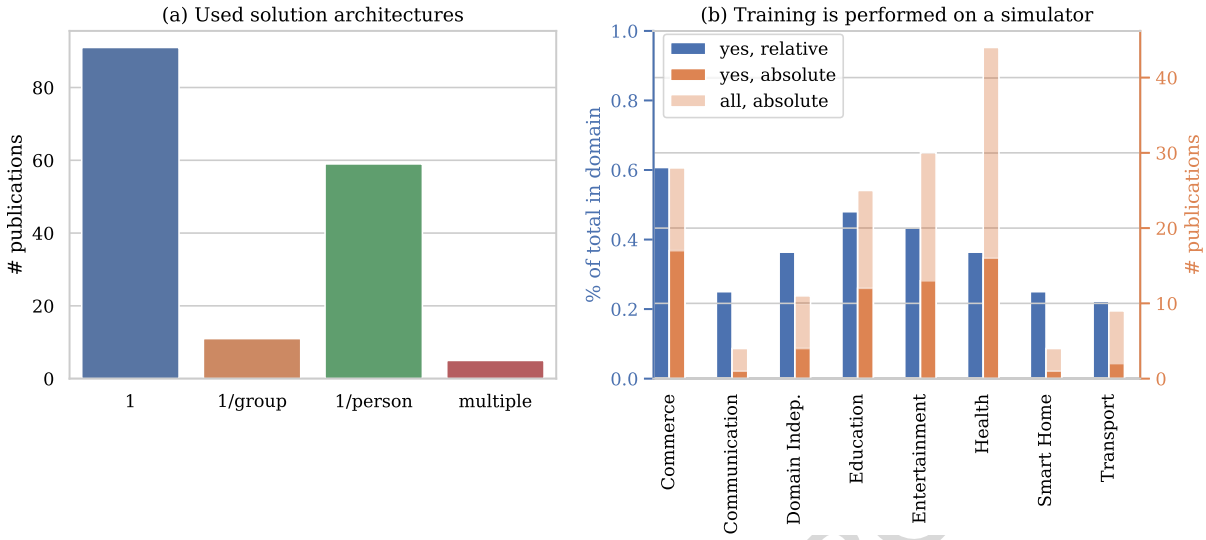

Fig. 9. Occurence of different solution architectures (a) and usage of simulators in training (b). For (a), publications that compare architectures are represented in the 'multiple' category.

Table 5
Number of models and the inclusion of user traits

Traits of users were used	Number of models			
	1	1/group	1/person	Multiple
In state representation	38	8	28	2
Other	5	0	9	3
Not used	48	3	22	0
Total	91	11	59	5

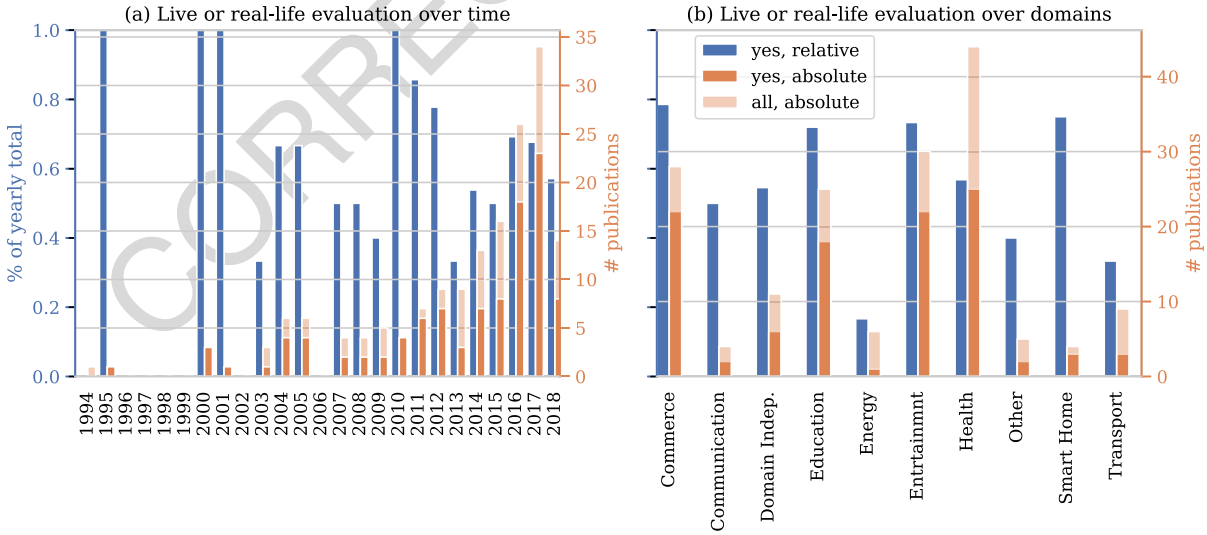

Fig. 10. Number of papers with a 'live' evaluation or evaluation using data on user responses to system behavior.

Table 6
Comparison of settings with realistic and other evaluation

	Real-world evaluation		Other evaluation	
	Count	% of column total	Count	% of column total
Total	104	100.0%	63	100.0%
Data on user responses to system behavior are available	57	54.8%	9	14.5%
Safety is mentioned as a concern in the article	14	13.5%	16	25.8%
Models of user responses to system behavior are available	21	20.2%	20	32.3%
Privacy is mentioned as a concern in the article	7	6.7%	2	3.2%
New interactions with users can be sampled with ease	60	57.7%	37	59.7%

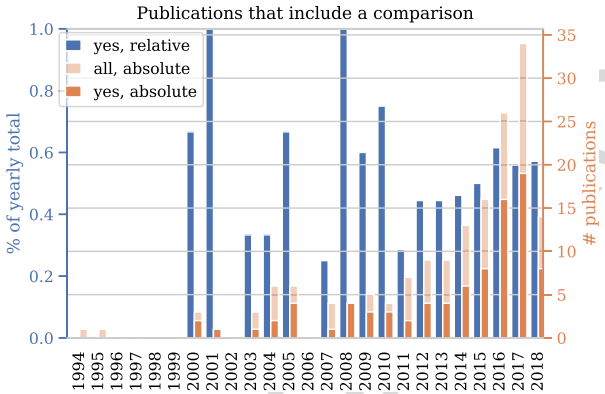

Fig. 11. Number of papers that include any comparison between solutions over time.

safety is mentioned as a concern. Relatively, studies that refrain from a realistic evaluation mention safety concerns almost twice as often as studies that do a realistic evaluation. The third biggest difference can be found in availability of user models. If a model is available, user responses can be simulated more easily. Privacy concerns are not mentioned frequently, so little can be said on its contribution to a lacking realistic evaluation. Finally and surprisingly, the ease of sampling interactions is comparable between studies with a realistic and without realistic evaluation.

Figure 11 describes how many studies include any of the comparisons in scope in this survey, that is: comparisons between solutions with and without personalization, comparisons between RL approaches and other approaches to personalization and comparisons between different RL algorithms. In the first years, no papers includes such a comparison. The period 2000-2010 contains relatively little studies in general and the absolute and relative numbers of studies with a comparison vary. From 2011 to 2018, the absolute number maintains it upward trend. The relative number follows this trend but flattens after 2016.

7. Discussion

The goal of this study was to give an overview and categorization of RL applications for personalization in different application domains which we addressed using a SLR on settings, solution architectures and evaluation strategies. The main result is the marked increase in studies that use RL for personaliza-

tion problems over time. Additionally, techniques are increasingly evaluated on real-life data. RL has proven a suitable paradigm for adaptation of systems to individual preferences using data.

Results further indicate that this development is driven by various techniques, which we list in no particular order. Firstly, techniques have been developed to estimate the performance of deploying a particular RL model prior to deployment. This helps in communicating risks and benefits of RL solutions with stakeholders and moves RL further into the realm of feasible technologies for high-impact application domains [200]. For single-step decision making problems, contextual bandit algorithms with theoretical bounds on decision-theoretic regret have become available. For multi-step decision making problems, methods that can estimate the performance of some policy based on data generated by another policy have been developed [37,90,204]. Secondly, advances in the field of deep learning have wholly or partly removed the need for feature engineering [53]. This may be especially challenging for sequential decision-making problems as different features may be of importance in different states encountered over time. Finally, research on safe exploration in RL has developed means to avoid harmful actions during exploratory phases of learning [64]. How any these techniques are best applied depends on setting. The collected data can be used to find suitable related work for any particular setting [46].

Since the field of RL for personalization is growing in size, we investigated whether methodological maturity is keeping pace. Results show that the growth in the *number* of studies with a real-life evaluation is not mirrored by growth of the *ratio* of studies with such an evaluation. Similarly, results show no increase in the relative number of studies with a comparison of approaches over time. These may be signs that the maturity of the field fails to keep pace with its growth. This is worrisome, since the advantages of RL over other approaches or between RL algorithms cannot be understood properly without such comparisons. Such comparisons benefit from standardized tasks. Developing standardized personalization datasets and simulation environments is an excellent opportunity for future research [87,112].

We found that algorithms presented in literature are reused infrequently. Although this phenomenon may be driven by various different underlying dynamics that cannot be untangled using our data, we propose some possible explanations here without particular order. Firstly, it might be the case that separate applications require tailored algorithms to the extend that these can only be used once. This raises the question on the scientific contribution of such a tailored algorithm and does not fit with the reuse of some well-established algorithms. Another explanation is that top-ranked venues prefer contributions that are theoretical or technical in nature, resulting in minor variations to well-known algorithms being presented as novel. Whether this is the case is out of scope for this research and forms an excellent avenue for future work. A final explanation for us to propose, is the myriad axes along which any RL algorithm can be identified, such as whether and where estimation is involved, which estimation technique is used and how domain knowledge is encoded in the algorithm. This may yield a large number of unique algorithms, constructed out of a relatively small set of core ideas in RL. An overview of these core ideas would be useful in understanding how individual algorithms relate to each other.

On top of algorithm reuse, we analyzed which RL algorithms were used most frequently. Generic and well-established (families of) algorithms such as Q-learning are the most popular. A notable entry in the top six most-used techniques is inverse reinforcement learning (IRL). Its frequent usage is surprising, as the only viable application area of IRL under a decade ago was robotics [97]. Personalization may be one of the other useful application areas of this branch of RL and many existing personalization challenges may still benefit from an IRL approach. Finally, we investigated how many RL models were included in the proposed solutions and found that the majority of studies resorts to using either one RL model in total or one RL model per user. Inspired by common practice of clustering in the related fields such as

e.g. recommender systems, we believe that there exists opportunities in pooling data of similar users and training RL models on the pooled data.

Besides these findings, we contribute a categorization of personalization settings in RL. This framework can be used to find related work based on the setting of a problem at hand. In designing such a framework, one has to balance specificity and usefulness of aspects in the framework. We take the aspect of 'safety' as an example: any application of RL will imply safety concerns at some level, but they are more prominent in some application areas. The framework intentionally includes a single ambiguous aspect to describe a broad range 'safety sensitivity levels' in order for it to suit its purpose of navigating literature. A possibility for future work is to extend the framework with other, more formal, aspects of problem setting such as those identified in [170].

Acknowledgements

The authors would like to thank Frank van Harmelen for useful feedback on the presented classification of personalization settings.

The authors declare that they have no conflict of interest.

Appendix A. Queries

```
TITLE-ABS-KEY(
("reinforcement learning" OR "contextual bandit") AND
("personalization" OR "personalized" OR "personal" OR
   "personalisation" OR "personalised" OR
"customization" OR "customized" OR "customised" OR "customised" OR
"individualized" OR "individualised" OR "tailored"))
```

Listing 1. Query for Scopus database

```
(((reinforcement learning) OR contextual bandit) AND
(personalization OR personalized OR personal OR personalisation OR
   personalised OR
customization OR customized OR customised OR customised OR
individualized OR individualised OR tailored))
```

Listing 2. Query for IEEE Xplore database command search

```
("reinforcement learning" OR "contextual bandit") AND
(personalization OR personalized OR personal OR personalisation OR
   personalised OR
customization OR customized OR customised OR customised OR
individualized OR individualised OR tailored)
```

Listing 3. Query for ACM DL database

```
reinforcement learning
(personalization | personalized | personal | personalisation | personalised |
customization | customized | customised | customised |
individualized | individualised | tailored)
```

Listing 4. First query for DBLP database

```
contextual bandit
(personalization | personalized | personal | personalisation | personalised |
customization | customized | customised | customised |
individualized | individualised | tailored)
```

Listing 5. Second query for DBLP database

```
allintitle: "reinforcement learning"
personalization OR personalized OR personal OR personalisation OR
   personalised OR
customization OR customized OR customised OR customised OR
individualized OR individualised OR tailored
```

Listing 6. First query for Google Scholar database

```
allintitle: "contextual bandit"
personalization OR personalized OR personal OR personalisation OR
   personalised OR
customization OR customized OR customised OR customised OR
individualized OR individualised OR tailored
```

Listing 7. Second query for Google Scholar database

Appendix B. Tabular view of data

Table 7
Table containing all included publications. The first column refers to the data items in Table 2

#	Value	Publications
1	n	[1,4,10,11,13,16,18–20,24–29,31,32,35,36,38,40–45,48,49,52,56,63,66,68,70,74,82,85,86,88,91,93,94,99,101,104,106–108,110,113,115–117,120,121,123–132,134,137–141,144,145,147,148,155–160,162,167,169,171,173–175,177,182–185,192–198,200,201,205–207,211,215,216,218,221–225,227,230–232,234–242]
	y	[3,6,8,33,34,51,57–62,72,73,80,83,103,109,114,119,142,150,151,151,161,168,176,180,183,208–210,213,214,217,226,228,229,233]
2	n	[3,6,10,13,24,28,36,38,40,45,49,51,57,59,72,73,103,109,114,126,129,131,142,147,148,150,159,162,183,213,214,217,226–228,233]
	y	[1,4,8,11,16,18–20,25–27,29,31–35,41–44,48,52,56,58,60–63,66,68,70,74,80,82,83,85,86,88,91,93,94,99,101,104,106–108,110,113,115–117,119–121,123–125,127,128,130,132,134,137–141,144,145,151,151,155–158,160,161,167–169,171,173–177,180,182–185,192–198,200,201,205–211,215,216,218,221–225,229–232,234–242]

Table 7
(Continued)

#	Value	Publications
3	n	[1,3,4,6,8,10,11,13,16,18,19,25–29,31,33–36,38,40,42,48,49,51,52,56–61,63,66,70,72–74,82,83,86,88,91,93,94,99,103,104,107–110,114,115,117,120,121,123–125,127–131,134,137–139,141,142,144,145,147,150,151,151,155,157–162,167,171,173–177,180,182,183,183–185,192–198,201,205,207,211,214–218,221–231,233–242]
	y	[20,24,32,41,43–45,62,68,80,85,101,106,113,116,119,126,132,140,148,156,168,169,200,206,208–210,213,232]
4	n	[3,4,6,8,10,11,13,16,18–20,24–29,31–36,38,40–45,48,49,51,52,56–63,66,68,70,72,74,80,82,83,85,86,91,93,94,99,101,103,104,106,108–110,113–117,119–121,123–132,134,137–142,144,145,147,148,150,151,151,155–162,167–169,173–177,180,182–185,192–198,200,201,205–208,210,211,213–216,218,221–239,241,242]
	y	[1,73,88,107,171,183,209,217,240]
5	n	[1,3,6,10,11,13,19,24–29,31,33–35,38,40,42,49,51,52,56,57,62,63,70,72,74,80,82,85,86,88,91,93,94,99,103,104,106–110,113–117,119,121,123,127–129,131,132,134,137–139,142,144,145,147,148,151,151,155,156,158–162,167–169,171,173–175,177,180,182,183,183,184,192,193,195–197,200,201,205–207,209,210,213–218,221–223,227,228,230,231,233–242]
	y	[4,8,16,18,20,32,36,41,43–45,48,58–61,66,68,73,83,101,120,124–126,130,140,141,150,157,176,185,194,198,208,211,224–226,229,232]
6	n	[1,3,4,8,10,13,16,18,24,26–28,31–33,38,40–42,44,45,48,49,51,52,56,57,59,62,66,68,70,80,82,83,85,86,91,93,94,99,101,104,106,109,116,117,119–121,123–126,128,130–132,138–141,144,145,147,150,151,156,157,159,162,167,169,171,175–177,182–185,193,205–208,210,211,214,218,221,222,230,232,234–238,242]
	y	[6,11,19,20,25,29,34–36,43,58,60,61,63,72–74,88,103,107,108,110,113–115,127,129,134,137,142,148,151,155,158,160,161,168,173,174,180,183,192,194–198,200,201,209,213,215–217,223–229,231,233,239–241]
7	n	[1,11,13,16,27–29,31,32,35,36,38,40,43,48,49,51,52,57,63,68,70,74,80,94,101,103,104,106,109,113,116,121,123,125,126,131,132,137,139,141,142,147,150,155–157,159,162,168,171,173,177,180,182,183,205–207,209,210,213,218,228,230,232,234–236]
	y	[3,4,6,8,10,18–20,24–26,33,34,41,42,44,45,56,58–62,66,72,73,82,83,85,86,88,91,93,99,107,108,110,114,115,117,119,120,124,127–130,134,138,140,144,145,148,151,151,158,160,161,167,169,174–176,183–185,192–198,200,201,208,211,214–217,221–227,229,231,233,237–242]
8	n	[20,25,28,31,34,36,56,62,68,72,88,107,114,115,119,121,134,144,148,159,167,192,195,201,210,215–218,222,229,231,238,240]
	y	[1,3,4,6,8,10,11,13,16,18,19,24,26,27,29,32,33,35,38,40–45,48,49,51,52,57–61,63,66,70,73,74,80,82,83,85,86,91,93,94,99,101,103,104,106,108–110,113,116,117,120,123–132,137–142,145,147,150,151,151,155–158,160–162,168,169,171,173–177,180,182,183,183–185,193,194,196–198,200,205–209,211,213,214,221,223–228,230,232–237,239,241,242]

Table 7
(Continued)

#	Value	Publications
10	1	[1,4,10,11,18,26–29,31,34,35,38,40,41,43,45,48,52,56,63,68,70,82,83,85,86,88,93,94,101,103,104,107,109,110,113,114,121,125–127,132,137,140,141,144,147,157–160,169,171,173,176,177,182–185,192–197,200,201,206–209,211,213–217,223–226,228,230,232,234,235,238,241,242]
	1/group	[16,42,108,115,117,123,124,155,218,221,236]
	1/person	[6,8,13,19,20,24,25,32,33,36,44,51,57–62,66,72–74,80,91,99,106,116,119,120,128–131,138,139,142,145,148,150,151,151,156,161,162,167,168,174,175,180,198,205,210,222,227,229,231,233,237,239]
	multiple	[3,49,134,183,240]
11	not used	[6,8,10,13,16,18,24,26,28,31,33–35,38,40,42–45,52,57,59,63,66,70,72,82,83,85,88,93,94,99,103,104,109,114,115,119,121,127,128,132,137,138,144,145,147,151,159,168,169,173,177,182,183,185,192,193,198,200,205,211,213,214,217,222,223,226,228,237,239,241]
	other	[3,19,25,73,74,130,131,134,150,156,158,161,195,197,230,232,240]
	state representation	[1,4,11,20,27,29,32,36,41,48,49,51,56,58,60–62,68,80,86,91,101,106–108,110,113,116,117,120,123–126,129,139–142,148,151,155,157,160,162,167,171,174–176,180,183,184,194,196,201,206–210,215,216,218,221,224,225,227,229,231,233–236,238,242]
12	batch	[1,34,35,40,49,51,56,58–61,73,88,101,103,108,116,121,123–126,129,140,141,150,157,158,162,180,183–185,192,193,196–198,200,201,206,215,216,223,229,230,234–236,238,240,242]
	n	[222]
	online	[6,8,18–20,24–26,29,32,33,36,41,45,48,57,62,66,68,70,72,80,86,91,93,94,99,106,107,109,110,113–115,119,120,128,130–132,137,138,142,144,145,148,151,151,161,167–169,171,173–176,182,194,195,208,210,211,213,217,224–228,233,237,239,241]
	other	[44,74,82,134,139,155,218,231]
	unknown	[3,4,10,11,13,16,27,28,31,38,42,43,52,63,83,85,104,117,127,147,156,159,160,177,183,205,207,209,214,221,232]
13	n	[1,3,6,11,19,26–29,31,33–36,40,43,49,52,63,74,88,93,94,101,103,104,107–110,113,114,117,121,123,125–129,132,137,140–142,144,145,147,151,155,157,158,160–162,171,173–175,177,180,182,183,196–198,200,201,207,209,213–215,221–226,228,230–233,235,238,239,241]
	y	[4,8,10,13,16,18,20,24,25,32,38,41,42,44,45,48,51,56–62,66,68,70,72,73,80,82,83,85,86,91,99,106,115,116,119,120,124,130,131,134,138,139,148,150,151,156,159,167–169,176,183–185,192–195,205,206,208,210,211,216–218,227,229,234,236,237,240,242]
14	n	[3,4,6,8,10,13,16,18,20,24–29,31–34,36,38,41,42,44,45,56–63,66,68,70,72–74,83,85,86,91,93,94,99,104,106,110,115–117,119,120,124,130–132,138,139,144,147,148,150,151,151,156,159,161,167–169,171,173,174,176,182,183,192,205,207–211,213,214,216,218,222,226–229,234,236–239,241]
	y	[1,11,19,35,40,43,48,49,51,52,80,82,88,101,103,107–109,113,114,121,123,125–129,134,137,140–142,145,155,157,158,160,162,175,177,180,183–185,193–198,200,201,206,215,217,221,223–225,230–233,235,240,242]

Table 7
(Continued)

#	Value	Publications
15	n	[1,3,8,10,11,13,16,18,19,24,25,27,28,31,32,34,35,38,40–45,48,49,51,52,56,57,59,62,66,68,70,73,80,83,85,88,91,99,101,103,107–109,113–117,120,121,123–131,134,137–142,145,147,148,150,151,155–159,162,167–169,175–177,180,183–185,192–195,197,200,201,205–211,214–216,218,221–225,229,230,232–236,238,240,242]
	y	[4,6,20,26,29,33,36,58,60,61,63,72,74,82,86,93,94,104,106,110,119,132,144,151,160,161,171,173,174,182,183,196,198,213,217,226–228,231,237,239,241]
16	n	[1,3,6,11,19,26–29,31,33–36,40,49,52,63,74,80,86,88,93,94,101,103,104,107,108,110,113,114,117,121,123,125–129,132,137,140–142,144,145,147,151,155,157,158,160–162,171,173–175,177,180,182,183,196–198,200,201,207,209,210,213–215,221–226,228,231–233,235,238,239,241]
	y	[4,8,10,13,16,18,20,24,25,32,38,41–45,48,51,56–62,66,68,70,72,73,82,83,85,91,99,106,109,115,116,119,120,124,130,131,134,138,139,148,150,151,156,159,167–169,176,183–185,192–195,205,206,208,211,216–218,227,229,230,234,236,237,240,242]
17	n	[3,4,6,8,10,13,16,18,20,24–29,31–36,38,41–45,56–63,66,68,70,72–74,83,85,86,91,93,94,99,104,106,109,110,115–117,119,120,124,130–132,137–139,144,147,148,150,151,151,156,159,161,167–169,171,173,174,176,182,183,183,184,192,198,205,207–211,213,214,216,218,222,226–230,234,237–239,241]
	y	[1,11,19,40,48,49,51,52,80,82,88,101,103,107,108,113,114,121,123,125–129,134,140–142,145,155,157,158,160,162,175,177,180,185,193–197,200,201,206,215,217,221,223–225,231–233,235,236,240,242]
18	n	[1,3,8,10,11,13,16,18,19,24,25,27,28,31,32,38,40–45,48,49,51,52,56,57,59,62,66,68,70,73,80,85,88,91,99,101,103,107,108,113–117,120,121,123–131,134,138–142,145,147,148,150,151,155–159,162,167,169,175–177,180,184,192–195,197,200,201,205–211,214–216,218,221–223,229,230,232–236,242]
	y	[4,6,20,26,29,33–36,58,60,61,63,72,74,82,83,86,93,94,104,106,109,110,119,132,137,144,151,160,161,168,171,173,174,182,183,183,185,196,198,213,217,224–228,231,237–241]
19	n	[1,3,8,18,19,24–29,31,35,36,40–43,45,51,52,56–58,60–62,68,73,80,82,85,86,88,91,93,94,103,104,108,110,113–117,119–121,123,127–132,137,138,141,142,144,147,148,150,151,155,156,158,161,162,167–169,171,173,176,177,182,184,185,192–194,200,201,205,207–211,214,216,218,221–223,226,228,229,234,235,237–240,242]
	y	[4,6,10,11,13,16,20,32–34,38,44,48,49,59,63,66,70,72,74,83,99,101,106,107,109,124–126,134,139,140,145,151,157,159,160,174,175,180,183,183,195–198,206,213,215,217,224,225,227,230–233,236,241]
20	n	[1,3,4,6,8,10,11,13,18,19,24–29,31–34,36,38,40–45,49,51,52,56–63,66,70,72–74,80,82,85,86,88,91,93,101,104,107,108,110,113,116,117,119–121,123,125–128,131,132,137–140,144,147,148,151,155,156,159,161,162,167–169,176,177,182,183,183,185,194,195,197,198,200,201,205,207–210,213–216,218,221–223,226,228–230,234,236,237,239–241]
	y	[16,20,35,48,68,83,94,99,103,106,109,114,115,124,129,130,134,141,142,145,150,151,157,158,160,171,173–175,180,184,192,193,196,206,211,217,224,225,227,231–233,235,238,242]

Table 7
(Continued)

#	Value	Publications
Domain	Commerce	[1,19,27,49,56,82,86,108,114,117,119,120,129,134,144,160,193,196,197,200,201,218,226,227,230, 233,235,238]
	Commu-nication	[43,99,104,171]
	Domain Independent	[20,25,36,126,138,139,148,194,224,225,229]
	Education	[29,31,35,52,63,74,80,83,93,94,103,115,121,123,124,147,161,162,176,177,184,185,215,216,223]
	Energy	[91,130,131,150,211,237]
	Entertainment	[4,10,18,34,38,58–61,70,72,88,107,109,127,128,137,151,158,159,167,175,183,183,192,195,217,228,231,240]
	Health	[3,6,8,11,28,40–42,44,45,48,51,66,68,73,101,106,113,116,125,132,140–142,155–157,168,169,173,180,205–209,213,222,232,234,236,239,241,242]
	Other	[24,26,174,210,214]
	Smart Home	[33,57,110,198]
	Transport	[13,16,32,62,85,145,151,182,221]

References

[1] N. Abe, N. Verma, C. Apte and R. Schroko, Cross channel optimized marketing by reinforcement learning, in: *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '04*, 2004. doi:10.1145/1014052.1016912.

[2] G. Abowd, A. Dey, P. Brown, N. Davies, M. Smith and P. Steggles, Towards a better understanding of context and context-awareness, in: *Handheld and Ubiquitous Computing*, Springer, 1999, p. 319. doi:10.1007/3-540-48157-5_29.

[3] S. Ahrndt, M. Lützenberger and S.M. Prochnow, Using personality models as prior knowledge to accelerate learning about stress-coping preferences: (demonstration), in: *AAMAS*, 2016. doi:10.5555/2936924.2937221.

[4] G. Andrade, G. Ramalho, H. Santana and V. Corruble, Automatic computer game balancing, in: *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems – AAMAS '05*, 2005. doi:10.1145/1082473.1082648.

[5] M.G. Aspinall and R.G. Hamermesh, Realizing the promise of personalized medicine, *Harvard Business Review* **85**(10) (2007), 108. https://hbr.org/2007/10/realizing-the-promise-of-personalized-medicine.

[6] A. Atrash and J. Pineau, A Bayesian reinforcement learning approach for customizing human–robot interfaces, in: *Proceedings of the 13th International Conference on Intelligent User Interfaces – IUI '09*, 2008. doi:10.1145/1502650.1502700.

[7] P. Auer, N. Cesa-Bianchi and P. Fischer, Finite-time analysis of the multiarmed bandit problem, *Machine Learning* **47**(2–3) (2002), 235–256. doi:10.1023/A:1013689704352.

[8] S. Ávila-Sansores, F. Orihuela-Espina and L. Enrique-Sucar, Patient tailored virtual rehabilitation, in: *Converging Clinical and Engineering Research on Neurorehabilitation*, Biosystems & Biorobotics, Vol. 1, 2013, pp. 879–883. doi:10.1007/978-3-642-34546-3_143.

[9] N.F. Awad and M.S. Krishnan, The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization, *MIS Quarterly* **30**(1) (2006), 13–28. doi:10.2307/25148715.

[10] N. Bagdure and B. Ambudkar, Reducing delay during vertical handover, in: *2015 International Conference on Computing Communication Control and Automation*, 2015. doi:10.1109/ICCUBEA.2015.44.

[11] A. Baniya, S. Herrmann, Q. Qiao and H. Lu, Adaptive interventions treatment modelling and regimen optimization using sequential multiple assignment randomized trials (SMART) and Q-learning, in: *IIE Annual Conference. Proceedings*, Institute of Industrial and Systems Engineers (IISE), 2017, pp. 1187–1192. https://pdfs.semanticscholar.org/858e/ffd10b711ad6c86eff9c32cdc0bc320a6e1a.pdf.

[12] A.G. Barto, P.S. Thomas and R.S. Sutton, Some recent applications of reinforcement learning, 2017. https://people.cs.umass.edu/~pthomas/papers/Barto2017.pdf.

[13] A.L.C. Bazzan, Synergies between evolutionary computation and multiagent reinforcement learning, in: *Proceedings of the Genetic and Evolutionary Computation Conference Companion – GECCO '17*, 2017. doi:10.1145/3067695.3075970.

[14] M.G. Bellemare, Y. Naddaf, J. Veness and M. Bowling, The arcade learning environment: An evaluation platform for general agents, *Journal of Artificial Intelligence Research* **47** (2013), 253–279. doi:10.1613/jair.3912.

[15] R.E. Bellman, *Adaptive Control Processes: A Guided Tour*, Vol. 2045, Princeton University Press, 2015. doi:10.1002/nav.3800080314.

[16] H. Bi, O.J. Akinwande and E. Gelenbe, Emergency navigation in confined spaces using dynamic grouping, in: *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies*, 2015. doi:10.1109/NGMAST.2015.12.

[17] G. Biegel and V. Cahill, A framework for developing mobile, context-aware applications, in: *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications, PerCom 2004*, IEEE, 2004, pp. 361–365. doi:10.1109/PERCOM.2004.1276875.

[18] A. Bodas, B. Upadhyay, C. Nadiger and S. Abdelhak, Reinforcement learning for game personalization on edge devices, in: *2018 International Conference on Information and Computer Technologies (ICICT)*, 2018. doi:10.1109/INFOCT.2018.8356853.

[19] D. Bouneffouf, A. Bouzeghoub and A.L. Gançarski, Hybrid-ϵ-greedy for mobile context-aware recommender system, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, Vol. 7301, 2012, pp. 468–479. doi:10.1007/978-3-642-30217-6_39.

[20] J. Bragg, Mausam and D.S. Weld, Optimal testing for crowd workers, in: *AAMAS*, 2016. doi:10.5555/2936924.2937066.

[21] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang and W. Zaremba, Openai gym, Preprint, arXiv:1606.01540, 2016.

[22] P. Brusilovski, A. Kobsa and W. Nejdl, *The Adaptive Web: Methods and Strategies of Web Personalization*, Vol. 4321, Springer, 2007. doi:10.1007/978-3-540-72079-9.

[23] D. Budgen and P. Brereton, Performing systematic literature reviews in software engineering, in: *Proceedings of the 28th International Conference on Software Engineering*, ACM, 2006, pp. 1051–1052. doi:10.1145/1134285.1134500.

[24] A.B. Buduru and S.S. Yau, An effective approach to continuous user authentication for touch screen smart devices, in: *2015 IEEE International Conference on Software Quality, Reliability and Security*, 2015. doi:10.1109/QRS.2015.40.
[25] I. Casanueva, T. Hain, H. Christensen, R. Marxer and P. Green, Knowledge transfer between speakers for personalised dialogue management, in: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 12–21. doi:10.18653/v1/W15-4603.
[26] A. Castro-Gonzalez, F. Amirabdollahian, D. Polani, M. Malfaz and M.A. Salichs, Robot self-preservation and adaptation to user preferences in game play, a preliminary study, in: *2011 IEEE International Conference on Robotics and Biomimetics*, 2011. doi:10.1109/ROBIO.2011.6181679.
[27] L. Cella, Modelling user behaviors with evolving users and catalogs of evolving items, in: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization – UMAP '17*, 2017. doi:10.1145/3099023.3102251.
[28] B. Chakraborty and S.A. Murphy, Dynamic treatment regimes, *Annual Review of Statistics and Its Application* **1**(1) (2014), 447–464. doi:10.1146/annurev-statistics-022513-115553.
[29] J. Chan and G. Nejat, A learning-based control architecture for an assistive robot providing social engagement during cognitively stimulating activities, in: *2011 IEEE International Conference on Robotics and Automation*, 2011. doi:10.1109/ICRA.2011.5980426.
[30] R.K. Chellappa and R.G. Sin, Personalization versus privacy: An empirical examination of the online consumer's dilemma, *Information Technology and Management* **6**(2–3) (2005), 181–202. doi:10.1007/s10799-005-5879-y.
[31] J. Chen and Z. Yang, A learning multi-agent system for personalized information filtering, in: *Proceedings of the 2003 Joint Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia*, 2003. doi:10.1109/ICICS.2003.1292790.
[32] X. Chen, Y. Zhai, C. Lu, J. Gong and G. Wang, A learning model for personalized adaptive cruise control, in: *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017. doi:10.1109/IVS.2017.7995748.
[33] Z. Cheng, Q. Zhao, F. Wang, Y. Jiang, L. Xia and J. Ding, Satisfaction based Q-learning for integrated lighting and blind control, *Energy and Buildings* **127** (2016), 43–55. doi:10.1016/j.enbuild.2016.05.067.
[34] C.-Y. Chi, R.T.-H. Tsai, J.-Y. Lai and J.Y. Hsu, A reinforcement learning approach to emotion-based automatic playlist generation, in: *2010 International Conference on Technologies and Applications of Artificial Intelligence*, 2010. doi:10.1109/TAAI.2010.21.
[35] M. Chi, K. VanLehn, D. Litman and P. Jordan, Inducing effective pedagogical strategies using learning context features, in: *International Conference on User Modeling, Adaptation, and Personalization*, Lecture Notes in Computer Science, Vol. 6075, 2010, pp. 147–158. doi:10.1007/978-3-642-13470-8_15.
[36] Y.-S. Chiang, T.-S. Chu, C.D. Lim, T.-Y. Wu, S.-H. Tseng and L.-C. Fu, Personalizing robot behavior for interruption in social human–robot interaction, in: *2014 IEEE International Workshop on Advanced Robotics and Its Social Impacts*, 2014. doi:10.1109/ARSO.2014.7020978.
[37] W. Chu, L. Li, L. Reyzin and R. Schapire, Contextual bandits with linear payoff functions, in: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 208–214. http://proceedings.mlr.press/v15/chu11a.
[38] M. Claeys, S. Latre, J. Famaey and F. De Turck, Design and evaluation of a self-learning HTTP adaptive video streaming client, *IEEE Communications Letters* **18**(4) (2014), 716–719. doi:10.1109/LCOMM.2014.020414.132649.
[39] G. Da Silveira, D. Borenstein and F.S. Fogliatto, Mass customization: Literature review and research directions, *International Journal of Production Economics* **72**(1) (2001), 1–13. doi:10.1016/S0925-5273(00)00079-7.
[40] M. Daltayanni, C. Wang and R. Akella, A fast interactive search system for healthcare services, in: *2012 Annual SRII Global Conference*, 2012. doi:10.1109/SRII.2012.65.
[41] E. Daskalaki, P. Diem and S.G. Mougiakakou, Personalized tuning of a reinforcement learning control algorithm for glucose regulation, in: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013. doi:10.1109/EMBC.2013.6610293.
[42] E. Daskalaki, P. Diem and S.G. Mougiakakou, An actor–critic based controller for glucose regulation in type 1 diabetes, *Computer Methods and Programs in Biomedicine* **109**(2) (2013), 116–125. doi:10.1016/j.cmpb.2012.03.002.
[43] E. Daskalaki, P. Diem and S.G. Mougiakakou, Model-free machine learning in biomedicine: Feasibility study in type 1 diabetes, *PLoS ONE* **11**(7) (2016), e0158722. doi:10.1371/journal.pone.0158722.
[44] M. De Paula, G.G. Acosta and E.C. Martínez, On-line policy learning and adaptation for real-time personalization of an artificial pancreas, *Expert Systems with Applications* **42**(4) (2015), 2234–2255. doi:10.1016/j.eswa.2014.10.038.
[45] M. De Paula, L.O. Ávila and E.C. Martínez, Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes, *Applied Soft Computing* **35** (2015), 310–332. doi:10.1016/j.asoc.2015.06.041.
[46] F. den Hengst, E. Grua, A. el Hassouni and M. Hoogendoorn, Release of the systematic literature review into reinforcement learning for personalization, Zenodo, 2020. doi:10.5281/zenodo.3627118.
[47] F. den Hengst, M. Hoogendoorn, F. van Harmelen and J. Bosman, Reinforcement learning for personalized dialogue management, in: *IEEE/WIC/ACM International Conference on Web Intelligence*, 2019, pp. 59–67. doi:10.1145/3350546.3352501.

[48] K. Deng, J. Pineau and S. Murphy, Active learning for personalizing treatment, in: *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2011. doi:10.1109/ADPRL.2011.5967348.
[49] A.A. Deshmukh, Ü. Dogan and C. Scott, Multi-task learning for contextual bandits, in: *NIPS*, 2017. doi:10.5555/3295222.3295238.
[50] Y. Duan, X. Chen, R. Houthooft, J. Schulman and P. Abbeel, Benchmarking deep reinforcement learning for continuous control, in: *International Conference on Machine Learning*, 2016, pp. 1329–1338. http://proceedings.mlr.press/v48/duan16.html.
[51] A. Durand and J. Pineau, Adaptive treatment allocation using sub-sampled Gaussian processes, in: *2015 AAAI Fall Symposium Series*, 2015. https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11671.
[52] M. El Fouki, N. Aknin and K.E. El Kadiri, Intelligent adapted e-learning system based on deep reinforcement learning, in: *Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems – ICCWCS'17*, 2017. doi:10.1145/3167486.3167574.
[53] A. El Hassouni, M. Hoogendoorn, A.E. Eiben, M. van Otterlo and V. Muhonen, End-to-end personalization of digital health interventions using raw sensor data with deep reinforcement learning: A comparative study in digital health interventions for behavior change, in: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, 2019, pp. 258–264. doi:10.1145/3350546.3352527.
[54] A. el Hassouni, M. Hoogendoorn, M. van Otterlo and E. Barbaro, Personalization of health interventions using cluster-based reinforcement learning, in: *International Conference on Principles and Practice of Multi-Agent Systems*, Springer, 2018, pp. 467–475. doi:10.1007/978-3-030-03098-8_31.
[55] H. Fan and M.S. Poole, What is personalization? Perspectives on the design and implementation of personalization in information systems, *Journal of Organizational Computing and Electronic Commerce* **16**(3–4) (2006), 179–202. doi:10.1207/s15327744joce1603&4_2.
[56] J. Feng, H. Li, M. Huang, S. Liu, W. Ou, Z. Wang and X. Zhu, Learning to collaborate, in: *Proceedings of the 2018 World Wide Web Conference on World Wide Web – WWW '18*, 2018. doi:10.1080/10919392.2006.9681199.
[57] B. Fernandez-Gauna and M. Grana, Recipe tuning by reinforcement learning in the SandS ecosystem, in: *2014 6th International Conference on Computational Aspects of Social Networks*, 2014. doi:10.1109/CASoN.2014.6920422.
[58] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, Exploiting reinforcement learning to profile users and personalize web pages, in: *2014 IEEE 38th International Computer Software and Applications Conference Workshops*, 2014. doi:10.1109/COMPSACW.2014.45.
[59] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, User centered and context dependent personalization through experiential transcoding, in: *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, 2014. doi:10.1109/CCNC.2014.6940520.
[60] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, Automatic web content personalization through reinforcement learning, *Journal of Systems and Software* **121** (2016), 157–169. doi:10.1016/j.jss.2016.02.008.
[61] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, On personalizing web content through reinforcement learning, *Universal Access in the Information Society* **16**(2) (2017), 395–410. doi:10.1007/s10209-016-0463-2.
[62] L. Fournier, Learning capabilities for improving automatic transmission control, in: *Proceedings of the Intelligent Vehicles '94 Symposium*, 1994. doi:10.1109/IVS.1994.639561.
[63] A.Y. Gao, W. Barendregt and G. Castellano, Personalised human–robot co-adaptation in instructional settings using reinforcement learning, in: *IVA Workshop on Persuasive Embodied Agents for Behavior Change: PEACH 2017*, August 27, Stockholm, Sweden, 2017. http://www.diva-portal.org/smash/get/diva2:1162389/FULLTEXT01.pdf.
[64] J. García and F. Fernández, A comprehensive survey on safe reinforcement learning, *Journal of Machine Learning Research* **16**(1) (2015), 1437–1480. http://jmlr.org/papers/v16/garcia15a.html.
[65] A. Garivier and E. Moulines, On upper-confidence bound policies for switching bandit problems, in: *International Conference on Algorithmic Learning Theory*, Springer, 2011, pp. 174–188. doi:10.1007/978-3-642-24412-4_16.
[66] A.E. Gaweda, Improving management of anemia in end stage renal disease using reinforcement learning, in: *2009 International Joint Conference on Neural Networks*, 2009. doi:10.1109/IJCNN.2009.5179004.
[67] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada and M.E. Brier, Individualization of pharmacological anemia management using reinforcement learning, *Neural Networks* **18**(5) (2005), 826–834. doi:10.1016/j.neunet.2005.06.020.
[68] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada and M.E. Brier, Incorporating prior knowledge into Q-learning for drug delivery individualization, in: *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*, 2005. doi:10.1109/ICMLA.2005.40.
[69] C. Gentile, S. Li and G. Zappella, Online clustering of bandits, in: *International Conference on Machine Learning*, 2014, pp. 757–765. http://proceedings.mlr.press/v32/gentile14.html.
[70] B.S. Ghahfarokhi and N. Movahhedinia, A personalized QoE-aware handover decision based on distributed reinforcement learning, *Wireless Networks* **19**(8) (2013), 1807–1828. doi:10.1007/s11276-013-0572-2.

[71] G.S. Ginsburg and J.J. McCarthy, Personalized medicine: Revolutionizing drug discovery and patient care, *Trends in Biotechnology* **19**(12) (2001), 491–496. doi:10.1016/S0167-7799(01)01814-5.
[72] D. Glowacka, T. Ruotsalo, K. Konuyshkova, K. Athukorala, S. Kaski and G. Jacucci, Directing exploratory search: reinforcement learning from user interactions with keywords, in: *Proceedings of the 2013 International Conference on Intelligent User Interfaces – IUI '13*, 2013. doi:10.1145/2449396.2449413.
[73] Y. Goldberg and M.R. Kosorok, Q-learning with censored data, *The Annals of Statistics* **40**(1) (2012), 529–560. doi:10.1214/12-AOS968.
[74] G. Gordon, S. Spaulding, J.K. Westlund, J.J. Lee, L. Plummer, M. Martinez, M. Das and C. Breazeal, Affective personalization of a social robot tutor for children's second language skills, in: *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. doi:10.5555/3016387.3016461.
[75] E.M. Grua and M. Hoogendoorn, Exploring clustering techniques for effective reinforcement learning based personalization for health and wellbeing, in: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2018, pp. 813–820. doi:10.1109/SSCI.2018.8628621.
[76] X. Guo, Y. Sun, Z. Yan and N. Wang, Privacy-personalization paradox in adoption of mobile health service: The mediating role of trust, in: *PACIS 2012 Proceedings*, 2012, p. 27. https://aisel.aisnet.org/pacis2012/27.
[77] M.A. Hamburg and F.S. Collins, The path to personalized medicine, *N. Engl. J. Med.* **2010**(363) (2010), 301–304. doi:10.1056/NEJMp1006304.
[78] A. Hans, D. Schneegaß, A.M. Schäfer and S. Udluft, Safe exploration for reinforcement learning., in: *ESANN*, 2008, pp. 143–148. http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2008-36.pdf.
[79] F.M. Harper, X. Li, Y. Chen and J.A. Konstan, An economic model of user rating in an online recommender system, in: *International Conference on User Modeling*, Lecture Notes in Computer Science, Vol. 3538, 2005, pp. 307–316. doi:10.1007/11527886_40.
[80] J. Hemminghaus and S. Kopp, Adaptive behavior generation for child–robot interaction, in: *Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction – HRI '18*, 2018. doi:10.1145/3173386.3176916.
[81] T. Hester and P. Stone, Learning and using models, in: *Reinforcement Learning*, M. Wiering and M. Van Otterlo, eds, Adaptation, Learning, and Optimization, Vol. 12, Springer, 2012, p. 120. doi:10.1007/978-3-642-27645-3_4.
[82] D.N. Hill, H. Nassif, Y. Liu, A. Iyer and S.V.N. Vishwanathan, An efficient bandit algorithm for realtime multivariate optimization, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '17*, 2017. doi:10.1145/3097983.3098184.
[83] T. Hiraoka, G. Neubig, S. Sakti, T. Toda and S. Nakamura, Learning cooperative persuasive dialogue policies using framing, *Speech Communication* **84** (2016), 83–96. doi:10.1016/j.specom.2016.09.002.
[84] L. Hood and M. Flores, A personal view on systems medicine and the emergence of proactive P4 medicine: Predictive, preventive, personalized and participatory, *New Biotechnology* **29**(6) (2012), 613–624. doi:10.1016/j.nbt.2012.03.004.
[85] Z. Huajun, Z. Jin, W. Rui and M. Tan, Multi-objective reinforcement learning algorithm and its application in drive system, in: *2008 34th Annual Conference of IEEE Industrial Electronics*, 2008. doi:10.1109/IECON.2008.4757965.
[86] S. Huang and F. Lin, Designing intelligent sales-agent for online selling, in: *Proceedings of the 7th International Conference on Electronic Commerce – ICEC '05*, 2005. doi:10.1145/1089551.1089605.
[87] E. Ie, C. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu and C. Boutilier, RecSim: A configurable simulation platform for recommender systems, Preprint, arXiv:1909.04847, 2019.
[88] S. Jaradat, N. Dokoohaki, M. Matskin and E. Ferrari, Trust and privacy correlations in social networks: A deep learning framework, in: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016. doi:10.1109/ASONAM.2016.7752236.
[89] G. Jawaheer, M. Szomszor and P. Kostkova, Comparison of implicit and explicit feedback from an online music recommendation service, in: *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 2010, ACM, pp. 47–51. doi:10.1145/1869446.1869453.
[90] N. Jiang and L. Li, Doubly robust off-policy value evaluation for reinforcement learning, in: *International Conference on Machine Learning*, 2016, pp. 652–661. http://proceedings.mlr.press/v48/jiang16.html.
[91] Z. Jin and Z. Huajun, Multi-objective reinforcement learning algorithm and its improved convergency method, in: *2011 6th IEEE Conference on Industrial Electronics and Applications*, 2011. doi:10.1109/ICIEA.2011.5976002.
[92] S. Junges, N. Jansen, C. Dehnert, U. Topcu and J.-P. Katoen, Safety-constrained reinforcement learning for MDPs, in: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, Springer, 2016, pp. 130–146. doi:10.1007/978-3-662-49674-9_8.
[93] A.A. Kardan and O.R.B. Speily, Smart lifelong learning system based on Q-learning, in: *2010 Seventh International Conference on Information Technology: New Generations*, 2010. doi:10.1109/ITNG.2010.140.
[94] I. Kastanis and M. Slater, Reinforcement learning utilizes proxemics, *ACM Transactions on Applied Perception* **9**(1) (2012), 1–15. doi:10.1145/2134203.2134206.

[95] S. Keizer, S. Rossignol, S. Chandramohan and O. Pietquin, User simulation in the development of statistical spoken dialogue systems, in: *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, Springer, 2012, pp. 39–74. doi:10.1007/978-1-4614-4803-7_4.
[96] M.K. Khribi, M. Jemni and O. Nasraoui, Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval, in: *Eighth IEEE International Conference on Advanced Learning Technologies, ICALT'08*, IEEE, 2008, pp. 241–245. doi:10.1109/ICALT.2008.198.
[97] J. Kober and J. Peters, Reinforcement learning in robotics: A survey, in: *Reinforcement Learning*, M. Wiering and M. Van Otterlo, eds, Adaptation, Learning, and Optimization, Vol. 12, Springer, 2012, pp. 596–597. doi:10.1007/978-3-642-27645-3_18.
[98] V.R. Konda and J.N. Tsitsiklis, Actor-critic algorithms, in: *Advances in Neural Information Processing Systems*, 2000, pp. 1008–1014. doi:10.5555/3009657.3009799.
[99] I. Koukoutsidis, A learning strategy for paging in mobile environments, in: *5th European Personal Mobile Communications Conference 2003*, 2003. doi:10.1049/cp:20030322.
[100] R. Kozierok and P. Maes, A learning interface agent for scheduling meetings, in: *Proceedings of the 1st International Conference on Intelligent User Interfaces*, ACM, 1993, pp. 81–88. doi:10.1145/169891.169908.
[101] E.F. Krakow, M. Hemmer, T. Wang, B. Logan, M. Arora, S. Spellman, D. Couriel, A. Alousi, J. Pidala, M. Last et al., Tools for the precision medicine era: How to develop highly personalized treatment recommendations from cohort and registry data using Q-learning, *American Journal of Epidemiology* **186**(2) (2017), 160–172. doi:10.1093/aje/kwx027.
[102] T.L. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics* **6**(1) (1985), 4–22. doi:10.1016/0196-8858(85)90002-8.
[103] A.S. Lan and R.G. Baraniuk, A contextual bandits framework for personalized learning action selection, in: *EDM*, 2016. http://www.educationaldatamining.org/EDM2016/proceedings/paper_18.pdf.
[104] G. Lee, S. Bauer, P. Faratin and J. Wroclawski, Learning user preferences for wireless services provisioning, in: *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2004*, 2004, pp. 480–487. doi:10.5555/1018409.1018782.
[105] O. Lemon, Conversational interfaces, in: *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, Springer, 2012, pp. 1–4. doi:10.1007/978-1-4614-4803-7.
[106] K. Li and M.Q.-H. Meng, Personalizing a service robot by learning human habits from behavioral footprints, *Engineering* **1**(1) (2015), 079. doi:10.15302/J-ENG-2015024.
[107] L. Li, W. Chu, J. Langford and R.E. Schapire, A contextual-bandit approach to personalized news article recommendation, in: *Proceedings of the 19th International Conference on World Wide Web*, ACM, 2010, pp. 661–670. doi:10.1145/1772690.1772758.
[108] Z. Li, J. Kiseleva, M. de Rijke and A. Grotov, Towards learning reward functions from user interactions, in: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval – ICTIR '17*, 2017. doi:10.1145/3121050.3121098.
[109] E. Liebman and P. Stone, DJ-MC: A reinforcement-learning agent for music playlist recommendation, in: *AAMAS*, 2015. doi:10.5555/2772879.2772954.
[110] J. Lim, H. Son, D. Lee and D. Lee, An MARL-based distributed learning scheme for capturing user preferences in a smart environment, in: *2017 IEEE International Conference on Services Computing (SCC)*, 2017. doi:10.1109/SCC.2017.24.
[111] L.-J. Lin, Self-improving reactive agents based on reinforcement learning, planning and teaching, *Machine Learning* **8**(3–4) (1992), 293–321. doi:10.1007/BF00992699.
[112] Q. Liu, B. Cui, Z. Wei, B. Peng, H. Huang, H. Deng, J. Hao, X. Huang and K.-F. Wong, Building personalized simulator for interactive search, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019, pp. 5109–5115. doi:10.24963/ijcai.2019/710.
[113] Y. Liu, B. Logan, N. Liu, Z. Xu, J. Tang and Y. Wang, Deep reinforcement learning for dynamic treatment regimes on medical registry data, in: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017. doi:10.1109/ICHI.2017.45.
[114] Llorente and S.E. Guerrero, Increasing retrieval quality in conversational recommenders, *IEEE Transactions on Knowledge and Data Engineering* **24**(10) (2012), 1876–1888. doi:10.1109/TKDE.2011.116.
[115] H.M.S. Lotfy, S.M.S. Khamis and M.M. Aboghazalah, Multi-agents and learning: Implications for webusage mining, *Journal of Advanced Research* **7**(2) (2016), 285–295. doi:10.1016/j.jare.2015.06.005.
[116] C. Lowery and A.A. Faisal, Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control, in: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013. doi:10.1109/NER.2013.6696208.
[117] O. Madani and D. DeCoste, Contextual recommender problems [extended abstract], in: *Proceedings of the 1st International Workshop on Utility-Based Data Mining – UBDM '05*, 2005. doi:10.1145/1089827.1089838.

[118] P. Maes and R. Kozierok, Learning interface agents, in: *AAAI*, Vol. 93, 1993, pp. 459–465. https://www.aaai.org/Papers/AAAI/1993/AAAI93-069.pdf.
[119] T. Mahmood, G. Mujtaba and A. Venturini, Dynamic personalization in conversational recommender systems, *Information Systems and e-Business Management* **12**(2) (2013), 213–238. doi:10.1007/s10257-013-0222-3.
[120] T. Mahmood and F. Ricci, Learning and adaptivity in interactive recommender systems, in: *Proceedings of the Ninth International Conference on Electronic Commerce – ICEC '07*, 2007. doi:10.1145/1282100.1282114.
[121] A. Malpani, B. Ravindran and H. Murthy, Personalized intelligent tutoring system using reinforcement learning, in: *FLAIRS Conference*, 2011. https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS11/paper/viewPaper/2597.
[122] U. Manber, A. Patel and J. Robison, Experience with personalization of Yahoo!, *Communications of the ACM* **43**(8) (2000), 35–39. doi:10.1145/345124.345136.
[123] I. Manickam, A.S. Lan and R.G. Baraniuk, Contextual multi-armed bandit algorithms for personalized learning action selection, in: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. doi:10.1109/ICASSP.2017.7953377.
[124] K.N. Martin and I. Arroyo, AgentX: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems, in: *International Conference on Intelligent Tutoring Systems*, Springer, 2004, pp. 564–572. doi:10.1007/978-3-540-30139-4_53.
[125] J.D. Martín-Guerrero, F. Gomez, E. Soria-Olivas, J. Schmidhuber, M. Climente-Martí and N.V. Jiménez-Torres, A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients, *Expert Systems with Applications* **36**(6) (2009), 9737–9742. doi:10.1016/j.eswa.2009.02.041.
[126] J.D. Martín-Guerrero, E. Soria-Olivas, M. Martínez-Sober, A.J. Serrrano-López, R. Magdalena-Benedito and J. Gómez-Sanchis, Use of reinforcement learning in two real applications, in: *Recent Advances in Reinforcement Learning*, 2008, pp. 191–204. doi:10.1007/978-3-540-89722-4_15.
[127] D. Massimo, M. Elahi and F. Ricci, Learning user preferences by observing user-items interactions in an IoT augmented space, in: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization – UMAP '17*, 2017. doi:10.1145/3099023.3099070.
[128] K. Masumitsu and T. Echigo, Video summarization using reinforcement learning in eigenspace, in: *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, 2000. doi:10.1109/ICIP.2000.899351.
[129] B.C. May, N. Korda, A. Lee and D.S. Leslie, Optimistic Bayesian sampling in contextual-bandit problems, *Journal of Machine Learning Research* **13** (2012), 2069–2106. http://jmlr.org/papers/v13/may12a.html.
[130] E. Mengelkamp, J. Gärttner and C. Weinhardt, Intelligent agent strategies for residential customers in local electricity markets, in: *Proceedings of the Ninth International Conference on Future Energy Systems – e-Energy '18*, 2018. doi:10.1145/3208903.3208907.
[131] E. Mengelkamp and C. Weinhardt, Clustering household preferences in local electricity markets, in: *Proceedings of the Ninth International Conference on Future Energy Systems – e-Energy '18*, 2018. doi:10.1145/3208903.3214348.
[132] N. Merkle and S. Zander, Agent-based assistance in ambient assisted living through reinforcement learning and semantic technologies, in: *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, Lecture Notes in Computer Science, Vol. 10574, 2017, pp. 180–188. doi:10.1007/978-3-319-69459-7_12.
[133] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, Playing atari with deep reinforcement learning, Preprint, arXiv:1312.5602, 2013.
[134] K. Mo, Y. Zhang, S. Li, J. Li and Q. Yang, Personalizing a dialogue system with transfer reinforcement learning, in: *AAAI*, 2018. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16104.
[135] D. Moher, A. Liberati, J. Tetzlaff and D.G. Altman, Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement, *Annals of Internal Medicine* **151**(4) (2009), 264–269. doi:10.7326/0003-4819-151-4-200908180-00135.
[136] T.M. Moldovan and P. Abbeel, Safe exploration in Markov decision processes, Preprint, arXiv:1205.4810, 2012.
[137] O. Moling, L. Baltrunas and F. Ricci, Optimal radio channel recommendations with explicit and implicit feedback, in: *Proceedings of the Sixth ACM Conference on Recommender Systems – RecSys '12*, 2012. doi:10.1145/2365952.2365971.
[138] A. Moon, T. Kang, H. Kim and H. Kim, A service recommendation using reinforcement learning for network-based robots in ubiquitous computing environments, in: *RO-MAN 2007 – The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 2007. doi:10.1109/ROMAN.2007.4415198.
[139] S. Narvekar, J. Sinapov and P. Stone, Autonomous task sequencing for customized curriculum design in reinforcement learning, in: *IJCAI*, 2017. doi:10.5555/3172077.3172241.
[140] S. Nemati, M.M. Ghassemi and G.D. Clifford, Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach, in: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016. doi:10.1109/EMBC.2016.7591355.
[141] D. Neumann, T. Mansi, L. Itu, B. Georgescu, E. Kayvanpour, F. Sedaghat-Hamedani, A. Amr, J. Haas, H. Katus, B. Meder et al., A self-taught artificial agent for multi-physics computational model personalization, *Medical Image Analysis* **34** (2016), 52–64. doi:10.1016/j.media.2016.04.003.

[142] D. Neumann, T. Mansi, L. Itu, B. Georgescu, E. Kayvanpour, F. Sedaghat-Hamedani, J. Haas, H. Katus, B. Meder, S. Steidl et al., Vito – A generic agent for multi-physics model personalization: Application to heart modeling, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 442–449. doi:10.1007/978-3-319-24571-3_53.

[143] D.W. Oard, J. Kim et al., Implicit feedback for recommender systems, in: *Proceedings of the AAAI Workshop on Recommender Systems*, AAAI Press, Menlo Park, CA, 1998, pp. 81–83. https://www.aaai.org/Papers/Workshops/1998/WS-98-08/WS98-08-021.pdf.

[144] D. Oh and C.L. Tan, Making better recommendations with online profiling agents, *AI Magazine* **26** (2004), 29–40. doi:10.1609/aimag.v26i3.1823.

[145] P. Ondruska and I. Posner, The route not taken: Driver-centric estimation of electric vehicle range, in: *Twenty-Fourth International Conference on Automated Planning and Scheduling*, 2014. https://www.aaai.org/ocs/index.php/ICAPS/ICAPS14/paper/viewPaper/7899.

[146] S.J. Pan and Q. Yang, A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering* **22**(10) (2010), 1345–1359. doi:10.1109/TKDE.2009.191.

[147] V. Pant, S. Bhasin and S. Jain, Self-learning system for personalized e-learning, in: *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*, 2017. doi:10.1109/ICETCCT.2017.8280344.

[148] P. Patompak, S. Jeong, I. Nilkhamhang and N.Y. Chong, Learning social relations for culture aware interaction, in: *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, 2017. doi:10.1109/URAI.2017.7992879.

[149] M. Pecka and T. Svoboda, Safe exploration techniques for reinforcement learning – An overview, in: *International Workshop on Modelling and Simulation for Autonomous Systems*, Springer, 2014, pp. 357–375. doi:10.1007/978-3-319-13823-7_31.

[150] B. Peng, Q. Jiao and T. Kurner, Angle of arrival estimation in dynamic indoor THz channels with Bayesian filter and reinforcement learning, in: *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016. doi:10.1109/EUSIPCO.2016.7760594.

[151] C. Peng and P. Vuorimaa, Automatic navigation among mobile DTV services, in: *ICEIS*, 2004. doi:10.5220/0002629501400145.

[152] C. Perera, A. Zaslavsky, P. Christen and D. Georgakopoulos, Context aware computing for the Internet of things: A survey, *IEEE Communications Surveys & Tutorials* **16**(1) (2014), 414–454. doi:10.1109/SURV.2013.042313.00197.

[153] T.J. Perkins and A.G. Barto, Lyapunov design for safe reinforcement learning, *Journal of Machine Learning Research* **3** (2002), 803–832. http://www.jmlr.org/papers/v3/perkins02a.html.

[154] B.J. Pine, B. Victor and A.C. Boynton, Making mass customization work, *Harvard Business Review* **71**(5) (1993), 108–111. https://hbr.org/1993/09/making-mass-customization-work.

[155] J. Pineau, M.G. Bellemare, A.J. Rush, A. Ghizaru and S.A. Murphy, Constructing evidence-based treatment strategies using methods from computer science, *Drug and Alcohol Dependence* **88** (2007), S52–S60. doi:10.1016/j.drugalcdep.2007.01.005.

[156] A. Pomprapa, S. Leonhardt and B.J.E. Misgeld, Optimal learning control of oxygen saturation using a policy iteration algorithm and a proof-of-concept in an interconnecting three-tank system, *Control Engineering Practice* **59** (2017), 194–203. doi:10.1016/j.conengprac.2016.07.014.

[157] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis and B.E. Engelhardt, A reinforcement learning approach to weaning of mechanical ventilation in intensive care units, Preprint, arXiv:1704.06300, 2017.

[158] M. Preda and D. Popescu, Personalized web recommendations: Supporting epistemic information about end-users, in: *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, 2005. doi:10.1109/WI.2005.115.

[159] F.D. Priscoli, L. Fogliati, A. Palo and A. Pietrabissa, Dynamic class of service mapping for quality of experience control in future networks, in: *WTC 2014; World Telecommunications Congress 2014*, VDE, 2014, pp. 1–6. https://www.vde-verlag.de/proceedings-de/453602012.html.

[160] Z. Qin, I. Rishabh and J. Carnahan, A scalable approach for periodical personalized recommendations, in: *Proceedings of the 10th ACM Conference on Recommender Systems – RecSys '16*, 2016. doi:10.1145/2959100.2959139.

[161] V.R. Raghuveer, B.K. Tripathy, T. Singh and S. Khanna, Reinforcement learning approach towards effective content recommendation in MOOC environments, in: *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)*, 2014. doi:10.1109/MITE.2014.7020289.

[162] E. Rennison, Personalized galaxies of information, in: *Companion of the ACM Conference on Human Factors in Computing Systems (CHI'95)*, 1995. doi:10.1145/223355.223409.

[163] P. Resnick and H.R. Varian, Recommender systems, *Communications of the ACM* **40**(3) (1997), 56–58. doi:10.1145/245108.245121.

[164] F. Ricci, L. Rokach and B. Shapira, Introduction to recommender systems handbook, in: *Recommender Systems Handbook*, Springer, 2011, pp. 14–17. doi:10.1007/978-0-387-85820-3.

[165] D. Riecken, Personalized views of personalization, *Communications of the ACM* **43**(8) (2000), 26–28. doi:10.1145/345124.345133.
[166] M. Riedmiller, Neural fitted Q iteration – First experiences with a data efficient neural reinforcement learning method, in: *European Conference on Machine Learning*, Springer, 2005, pp. 317–328. doi:10.1007/11564096_32.
[167] H. Ritschel and E. André, Real-time robot personality adaptation based on reinforcement learning and social signals, in: *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human–Robot Interaction – HRI '17*, 2017. doi:10.1145/3029798.3038381.
[168] I. Rivas-Blanco, C. Lopez-Casado, C.J. Perez-del-Pulgar, F. Garcia-Vacas, J.C. Fraile and V.F. Munoz, Smart cable-driven camera robotic assistant, *IEEE Transactions on Human–Machine Systems* **48**(2) (2018), 183–196. doi:10.1109/THMS.2017.2767286.
[169] M. Rudary, S. Singh and M.E. Pollack, Adaptive cognitive orthotics: Combining reinforcement learning and constraint-based temporal reasoning, in: *Proceedings of the Twenty-First International Conference on Machine Learning*, ACM, 2004, p. 91. doi:10.1145/2959100.2959139.
[170] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall, Egnlewood Cliffs, NJ, 1995. ISBN 0136042597.
[171] S. Saha and R. Quazi, Emotion-driven learning agent for setting rich presence in mobile telephony, in: *2008 11th International Conference on Computer and Information Technology*, 2008. doi:10.1109/ICCITECHN.2008.4803023.
[172] J.B. Schafer, D. Frankowski, J. Herlocker and S. Sen, Collaborative filtering recommender systems, in: *The Adaptive Web*, Springer, 2007, pp. 291–324. doi:10.1007/978-3-540-72079-9_9.
[173] Y.A. Sekhavat, MPRL: Multiple-periodic reinforcement learning for difficulty adjustment in rehabilitation games, in: *2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH)*, 2017. doi:10.1109/SeGAH.2017.7939260.
[174] Y.-W. Seo and B.-T. Zhang, A reinforcement learning agent for personalized information filtering, in: *Proceedings of the 5th International Conference on Intelligent User Interfaces*, ACM, 2000, pp. 248–251. doi:10.1145/325737.325859.
[175] Y.-W. Seo and B.-T. Zhang, Learning user's preferences by analyzing web-browsing behaviors, in: *Proceedings of the Fourth International Conference on Autonomous Agents*, ACM, 2000, pp. 381–387. doi:10.1145/336595.337546.
[176] D. Shawky and A. Badawi, A reinforcement learning-based adaptive learning system, in: *Advances in Intelligent Systems and Computing*, 2018, pp. 221–231. doi:10.1007/978-3-319-74690-6_22.
[177] S. Shen and M. Chi, Reinforcement learning: the Sooner the Better, or the Later the Better?, in: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization – UMAP '16*, 2016. doi:10.1145/2930238.2930247.
[178] S.M. Shortreed, E. Laber, D.J. Lizotte, T.S. Stroup, J. Pineau and S.A. Murphy, Informing sequential clinical decision-making through reinforcement learning: An empirical study, *Machine Learning* **84**(1–2) (2011), 109–136. doi:10.1007/s10994-010-5229-0.
[179] G.E. Simon and R.H. Perlis, Personalized medicine for depression: Can we match patients with treatments?, *American Journal of Psychiatry* **167**(12) (2010), 1445–1455. doi:10.1176/appi.ajp.2010.09111680.
[180] L. Song, W. Hsu, J. Xu and M. van der Schaar, Using contextual learning to improve diagnostic accuracy: Application in breast cancer screening, *IEEE Journal of Biomedical and Health Informatics* **20**(3) (2016), 902–914. doi:10.1109/JBHI.2015.2414934.
[181] N. Sprague and D. Ballard, Multiple-goal reinforcement learning with modular sarsa (0), 2003. doi:10.5555/1630659.1630892.
[182] A.R. Srinivasan and S. Chakraborty, Path planning with user route preference – A reward surface approximation approach using orthogonal Legendre polynomials, in: *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, 2016. doi:10.1109/COASE.2016.7743527.
[183] A. Srivihok and P. Sukonmanee, Intelligent agent for e-tourism: Personalization travel support agent using reinforcement learning, in: *WWW 2005*, 2005. http://ceur-ws.org/Vol-143/paper12.pdf.
[184] P. Su, Y.-B. Wang, T. Yu and L. Lee, A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning, in: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013. doi:10.1109/ICASSP.2013.6639266.
[185] P.-H. Su, C.-H. Wu and L.-S. Lee, A recursive dialogue game for personalized computer-aided pronunciation training, in: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014. doi:10.1109/TASLP.2014.2375572.
[186] R.S. Sutton, Integrated architectures for learning, planning, and reacting based on approximating dynamic programming, in: *Machine Learning Proceedings 1990*, Elsevier, 1990, pp. 216–224. doi:10.1016/B978-1-55860-141-3.50030-4.
[187] R.S. Sutton, Generalization in reinforcement learning: Successful examples using sparse coarse coding, in: *Advances in Neural Information Processing Systems*, 1996, pp. 1038–1044. doi:10.5555/2998828.2998974.
[188] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2018. ISBN 9780262193986.

[189] R.S. Sutton, D.A. McAllester, S.P. Singh and Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in: *Advances in Neural Information Processing Systems*, 2000, pp. 1057–1063. doi:10.5555/3009657.3009806.
[190] C. Szepesvári, Algorithms for reinforcement learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning* **4**(1) (2010), 1–103. doi:10.2200/S00268ED1V01Y201005AIM009.
[191] S.A. Tabatabaei, M. Hoogendoorn and A. van Halteren, Narrowing reinforcement learning: Overcoming the cold start problem for personalized health interventions, in: *International Conference on Principles and Practice of Multi-Agent Systems*, Springer, 2018, pp. 312–327. doi:10.1007/978-3-030-03098-8_19.
[192] N. Taghipour and A. Kardan, A hybrid web recommender system based on Q-learning, in: *Proceedings of the 2008 ACM Symposium on Applied Computing – SAC '08*, 2008. doi:10.1145/1363686.1363954.
[193] N. Taghipour, A. Kardan and S.S. Ghidary, Usage-based web recommendations, in: *Proceedings of the 2007 ACM Conference on Recommender Systems – RecSys '07*, 2007. doi:10.1145/1297231.1297250.
[194] L. Tang, Y. Jiang, L. Li and T. Li, Ensemble contextual bandits for personalized recommendation, in: *Proceedings of the 8th ACM Conference on Recommender Systems – RecSys '14*, 2014. doi:10.1145/2645710.2645732.
[195] L. Tang, Y. Jiang, L. Li, C. Zeng and T. Li, Personalized recommendation via parameter-free contextual bandits, in: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '15*, 2015. doi:10.1145/2766462.2767707.
[196] L. Tang, R. Rosales, A. Singh and D. Agarwal, Automatic ad format selection via contextual bandits, in: *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management – CIKM '13*, 2013. doi:10.1145/2505515.2514700.
[197] M. Tavakol and U. Brefeld, A unified contextual bandit framework for long- and short-term recommendations, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, Vol. 10535, 2017, pp. 269–284. doi:10.1007/978-3-319-71246-8_17.
[198] B. Tegelund, H. Son and D. Lee, A task-oriented service personalization scheme for smart environments using reinforcement learning, in: *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 2016. doi:10.1109/PERCOMW.2016.7457110.
[199] G. Tesauro, Temporal difference learning and TD-Gammon, *Communications of the ACM* **38**(3) (1995), 58–68. doi:10.1145/203330.203343.
[200] G. Theocharous, P.S. Thomas and M. Ghavamzadeh, Personalized ad recommendation systems for life-time value optimization with guarantees, in: *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. doi:10.5555/2832415.2832500.
[201] G. Theocharous, P.S. Thomas and M. Ghavamzadeh, Ad recommendation systems for life-time value optimization, in: *Proceedings of the 24th International Conference on World Wide Web – WWW '15 Companion*, 2015. doi:10.1145/2740908.2741998.
[202] P. Thomas and E. Brunskill, Data-efficient off-policy policy evaluation for reinforcement learning, in: *International Conference on Machine Learning*, 2016, pp. 2139–2148. http://proceedings.mlr.press/v48/thomasa16.html.
[203] P.S. Thomas, Safe reinforcement learning, 2015. https://scholarworks.umass.edu/dissertations_2/514.
[204] P.S. Thomas, G. Theocharous and M. Ghavamzadeh, High-confidence off-policy evaluation, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. doi:10.5555/2888116.2888134.
[205] S. Triki and C. Hanachi, A self-adaptive system for improving autonomy and public spaces accessibility for elderly, in: *Smart Innovation, Systems and Technologies*, 2017, pp. 53–66. doi:10.1007/978-3-319-59394-4_6.
[206] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R.K. Ten Haken and I.E. Naqa, Deep reinforcement learning for automated radiation adaptation in lung cancer, *Medical Physics* **44**(12) (2017), 6690–6705. doi:10.1002/mp.12625.
[207] K. Tsiakas, C. Abellanoza and F. Makedon, Interactive learning and adaptation for robot assisted therapy for people with dementia, in: *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments – PETRA '16*, 2016. doi:10.1145/2910674.2935849.
[208] K. Tsiakas, M. Huber and F. Makedon, A multimodal adaptive session manager for physical rehabilitation exercising, in: *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments – PETRA '15*, 2015. doi:10.1145/2769493.2769507.
[209] K. Tsiakas, M. Papakostas, B. Chebaa, D. Ebert, V. Karkaletsis and F. Makedon, An interactive learning and adaptation framework for adaptive robot assisted therapy, in: *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments – PETRA '16*, 2016. doi:10.1145/2910674.2935857.
[210] K. Tsiakas, M. Papakostas, M. Theofanidis, M. Bell, R. Mihalcea, S. Wang, M. Burzo and F. Makedon, An interactive multisensing framework for personalized human robot collaboration and assistive training using reinforcement learning, in: *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments – PETRA '17*, 2017. doi:10.1145/3056540.3076191.
[211] D. Urieli and P. Stone, TacTex'13: A champion adaptive power trading agent, in: *AAMAS*, 2014. doi:10.5555/2615731.2617516.

[212] H. Van Hasselt, A. Guez and D. Silver, Deep reinforcement learning with double q-learning, in: *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. doi:10.5555/3016100.3016191.
[213] G. Vasan and P.M. Pilarski, Learning from demonstration: Teaching a myoelectric prosthesis with an intact limb via reinforcement learning, in: *2017 International Conference on Rehabilitation Robotics (ICORR)*, 2017. doi:10.1109/ICORR.2017.8009453.
[214] L. Wang, Y. Gao, C. Cao and L. Wang, Towards a general supporting framework for self-adaptive software systems, in: *2012 IEEE 36th Annual Computer Software and Applications Conference Workshops*, 2012. doi:10.1109/COMPSACW.2012.38.
[215] P. Wang, J. Rowe, B. Mott and J. Lester, Decomposing drama management in educational interactive narrative: A modular reinforcement learning approach, in: *International Conference on Interactive Digital Storytelling*, Lecture Notes in Computer Science, Vol. 10045, 2016, pp. 270–282. doi:10.1007/978-3-319-48279-8_24.
[216] P. Wang, J.P. Rowe, W. Min, B.W. Mott and J.C. Lester, Interactive narrative personalization with deep reinforcement learning, in: *IJCAI*, 2017. doi:10.5555/3172077.3172427.
[217] X. Wang, Y. Wang, D. Hsu and Y. Wang, Exploration in interactive personalized music recommendation: A reinforcement learning approach, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **11**(1) (2014), 7. doi:10.1145/2648583.
[218] X. Wang, M. Zhang, F. Ren and T. Ito, GongBroker: A broker model for power trading in smart grid markets, in: *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2015. doi:10.1109/WI-IAT.2015.108.
[219] C.J. Watkins and P. Dayan, Q-learning, *Machine Learning* **8**(3–4) (1992), 279–292. doi:10.1007/BF00992698.
[220] M. Wiering and M. Van Otterlo, Reinforcement learning, in: *Reinforcement Learning*, M. Wiering and M. Van Otterlo, eds, Adaptation, Learning, and Optimization, Vol. 12, Springer, 2012. doi:10.1007/978-3-642-27645-3.
[221] G. Wu, Y. Ding, Y. Li, J. Luo, F. Zhang and J. Fu, Data-driven inverse learning of passenger preferences in urban public transits, in: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017. doi:10.1109/CDC.2017.8264410.
[222] J. Xu, T. Xing and M. van der Schaar, Personalized course sequence recommendations, *IEEE Transactions on Signal Processing* **64**(20) (2016), 5340–5352. doi:10.1109/TSP.2016.2595495.
[223] M. Yang, Q. Qu, K. Lei, J. Zhu, Z. Zhao, X. Chen and J.Z. Huang, Investigating deep reinforcement learning techniques in personalized dialogue generation, in: *Proceedings of the 2018 SIAM International Conference on Data Mining*, SIAM, 2018, pp. 630–638. doi:10.1137/1.9781611975321.71.
[224] M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen and J. Zhu, Personalized response generation by dual-learning based domain adaptation, *Neural Networks* **103** (2018), 72–82. doi:10.1016/j.neunet.2018.03.009.
[225] M. Yang, Z. Zhao, W. Zhao, X. Chen, J. Zhu, L. Zhou and Z. Cao, Personalized response generation via domain adaptation, in: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '17*, 2017. doi:10.1145/3077136.3080706.
[226] S.-T. Yuan, A personalized and integrative comparison-shopping engine and its applications, *Decision Support Systems* **34**(2) (2003), 139–156. doi:10.1016/S0167-9236(02)00077-5.
[227] Y. Yue, S.A. Hong and C. Guestrin, Hierarchical exploration for accelerating contextual bandits, in: *Proceedings of the 29th International Coference on International Conference on Machine Learning*, Omnipress, 2012, pp. 979–986. doi:10.5555/3042573.3042700.
[228] S. Zaidenberg and P. Reignier, Reinforcement learning of user preferences for a ubiquitous personal assistant, in: *Advances in Reinforcement Learning*, IntechOpen, 2011. doi:10.5772/13723.
[229] S. Zaidenberg, P. Reignier and J.L. Crowley, Reinforcement learning of context models for a ubiquitous personal assistant, in: *3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008*, 2008, pp. 254–264. doi:10.1007/978-3-540-85867-6_30.
[230] C. Zeng, Q. Wang, S. Mokhtari and T. Li, Online context-aware recommendation with time varying multi-armed bandit, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16*, 2016. doi:10.1145/2939672.2939878.
[231] B.-T. Zhang and Y.-W. Seo, Personalized web-document filtering using reinforcement learning, *Applied Artificial Intelligence* **15**(7) (2001), 665–685. doi:10.1080/088395101750363993.
[232] Y. Zhang, R. Chen, J. Tang, W.F. Stewart and J. Sun, LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '17*, 2017. doi:10.1145/3097983.3098109.
[233] T. Zhao and I. King, Locality-sensitive linear bandit model for online social recommendation, in: *International Conference on Neural Information Processing*, Lecture Notes in Computer Science, Vol. 9947, 2016, pp. 80–90. doi:10.1007/978-3-319-46687-3_9.
[234] Y. Zhao, M.R. Kosorok and D. Zeng, Reinforcement learning design for cancer clinical trials, *Statistics in Medicine* **28**(26) (2009), 3294–3315. doi:10.1002/sim.3720.

[235] Y. Zhao, S. Wang, Y. Zou, J. Ng and T. Ng, Automatically learning user preferences for personalized service composition, in: *2017 IEEE International Conference on Web Services (ICWS)*, 2017. doi:10.1109/ICWS.2017.93.
[236] Y. Zhao, D. Zeng, M.A. Socinski and M.R. Kosorok, Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer, *Biometrics* **67**(4) (2011), 1422–1433. doi:10.1111/j.1541-0420.2011.01572.x.
[237] Y. Zhao, Q. Zhao, L. Xia, Z. Cheng, F. Wang and F. Song, A unified control framework of HVAC system for thermal and acoustic comforts in office building, in: *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, 2013. doi:10.1109/CoASE.2013.6653964.
[238] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N.J. Yuan, X. Xie and Z. Li, DRN: A deep reinforcement learning framework for news recommendation, in: *Proceedings of the 2018 World Wide Web Conference on World Wide Web – WWW '18*, 2018. doi:10.1145/3178876.3185994.
[239] H. Zheng and J. Jumadinova, OWLS: Observational wireless life-enhancing system (extended abstract), in: *AAMAS*, 2016. doi:10.5555/2936924.2937192.
[240] L. Zhou and E. Brunskill, Latent contextual bandits and their application to personalized recommendations for new users, in: *IJCAI*, 2016. doi:10.5555/3061053.3061129.
[241] M. Zhou, Y.D. Mintz, Y. Fukuoka, K.Y. Goldberg, E. Flowers, P. Kaminsky, A. Castillejo and A. Aswani, Personalizing mobile fitness apps using reinforcement learning, in: *IUI Workshops*, 2018. http://ceur-ws.org/Vol-2068/humanize7.pdf.
[242] R. Zhu, Y.-Q. Zhao, G. Chen, S. Ma and H. Zhao, Greedy outcome weighted tree learning of optimal personalized treatment rules, *Biometrics* **73**(2) (2016), 391–400. doi:10.1111/biom.12593.