Data Science 0 (2020) 0
IOS Press

Reinforcement learning for personalization: a systematic literature review

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

[a] *Dept. of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV, Amsterdam The Netherlands*
E-mails: f.den.hengst@vu.nl, e.m.grua@vu.nl, a.el.hassouni@vu.nl, m.hoogendoorn@vu.nl; ORCIDs: https://orcid.org/0000-0002-2092-9904, https://orcid.org/0000-0002-5471-4338, https://orcid.org/000-0003-0919-8861, https://orcid.org/0000-0003-3356-3574

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

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 [1, 2], human-computer interaction [3, 4], product, news, music and video recommendations [5–7] and even manufacturing [8, 9]. 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* [10].

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's[1] digital video delivery mechanism includes tracking of views and ratings. These ease the gratification of diverse entertainment needs as they enable Netflix

[*]Corresponding author. E-mail: f.den.hengst@vu.nl.
[**]Authors contributed equally.
[1]https://www.netflix.com

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 [11, 12].

RL has seen successful applications to personalization in a wide variety of domains. Some of the earliest work, such as [13], [14] and [15] focused on web services. More recently, [16] 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, [17] 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 [18]. A more recent example in this domain, [19], 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 [20–23]. 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.

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 4 details the SLR protocol, results and analysis are described in Section 5. All data collected has been made available digitally [24]. Finally, we conclude with current trends challenges in Section 6.

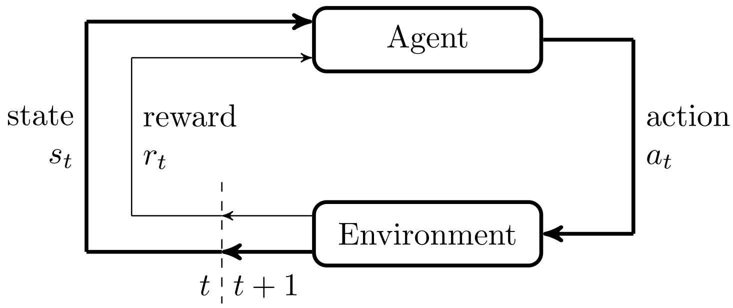

Fig. 1. The agent-environment in RL for personalization from [25].

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 Figure 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. We refer the reader to [25], [26] and [27] for such an 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, $T : S \times A \times S \rightarrow [0,1]$ a probabilistic transition function, $R : S \times A \rightarrow \mathbb{R}$ 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 $tr : \langle s^0, a^0, r^0, s^1, \ldots, a^{T-1}, r^{T-1}, s^T\rangle$. 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 $\Pi : S \times A \rightarrow [0,1]$ that maximizes 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 $\mathbb{E}$ 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:

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

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:

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

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 consider a set of n users $U \in \{u_1, \ldots, u_n\}$. In some settings, users can be described using a function that returns a vector representation of the l features that characterize a user $\phi : U \to \langle \phi_1(U), \ldots, \phi_l(U) \rangle$. A first way to adapt software systems to an individual users' needs is to define a separate MDP and corresponding RL agent for each user. All of the concepts introduced before can be indexed with subscript i for user u_i and the overall goal becomes to find an optimal policy per user: $\{\pi_1^*, \ldots, \pi_n^*\}$. In the case of approximations as in Eq. (5), these are made per user based on data set D_i with trajectories only involving that user. The benefit of isolated MDPs is that differences between T_i and T_j or between R_i and R_j for users $u_i, u_j, u_i \neq u_j$ 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 and every agent has to relearn the task at hand for each user individually. This may require a substantial number of experiences per user 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.

An alternative approach to finding per-user optimal policies $\{\pi_1^*, \ldots, \pi_n^*\}$ is to alter the state space S to include information about the user on top of information about the task at hand and then learn a single π^* for all users [28]. A natural extension of the state space S to S' is to add the feature vector representation of users $\phi(U)$ to S. This approach can be valuable when it is unclear which experiences of users $u_j \neq u_i$ should be used in determining π_i^*, i.e. when no suitable M can be defined upfront. Conceptually, finding $\pi^*(s')$ where $s' \in S'$ 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 [29]. 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 S and S' 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 pi^* per user. It is motivated by the idea that trajectories of similar users $u_j \neq u_i$ may prove useful in estimating $\mathbb{E}_{\pi_i}[G]$. One such an approach is based on clustering similar users [18, 30–32]. It requires $q \leqslant o$ groups $G \in \{g_1, \ldots, g_q\}$ and a mapping function $M : U \rightarrow G$. One MDP and RL agent are defined for every g_p to interact with users $u_i, u_j, M(u_i) = M(u_j) = g_p$. 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 u_i, u_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 for all users. A related approach similarly uses experiences D_j of other users $u_j \neq u_i$ but still aims to find per-user 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 $M : U \rightarrow G$ for the pooling approach and note that it can be rewritten to $M : U \times U \rightarrow \{0, 1\}$. The weighting scheme, now, is a generalization where $M : U \times U \rightarrow \mathbb{R}$. Finding a suitable M can be challenging in itself and depends on the availability of user features, trajectories and the task at hand. Typical strategies are to define M in terms of similarity of feature representations $[\phi(u_i), \phi(u_j)]$ and 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 M is suitable. If either of these assumptions is not met, pooling data may result in a policy that is suboptimal for both u_i and u_j. This phenomenon is typically referred to as the *negative transfer problem* [33].

3. A classification of personalization settings

Personalization has many different definitions [10, 34, 35]. We adopt the definition proposed in [10] 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 4.4 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 [36]. 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 [36, 37]. Besides suitability, we consider safety of system behavior. Unaltered RL algorithms use trial-and-error style exploration to optimize their behavior. If safety is a significant concern in the systems' application domain, specifically designed safety-aware RL techniques may be required, see [38] and [39] 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 [40]. Models can also be used to interact with the RL agent in simulation. Secondly, upfront knowledge may be available in

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

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 [41, 42].

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 [43]. 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 [7, 44, 45] and physical state [46, 47], both of which may be hard to measure accurately [48–50].

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.

4. 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' [51]. PRISMA is a standard for reporting on SLRs and details eligibility criteria, article collection, screening process, data extraction and data synthesis [52]. 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.

4.1. 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 3. 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.

4.2. Search strategy

Figure 2 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. 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 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.

4.3. Screening process

In the screening process, all query results are tested against the inclusion criteria in two phases. In the first phase, studies are assessed based on keywords, abstract and title. For this phase, the 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. Data on studies in this phase were copied to a new spreadsheet to which columns were added for all data items. 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.

4.4. 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 3. 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

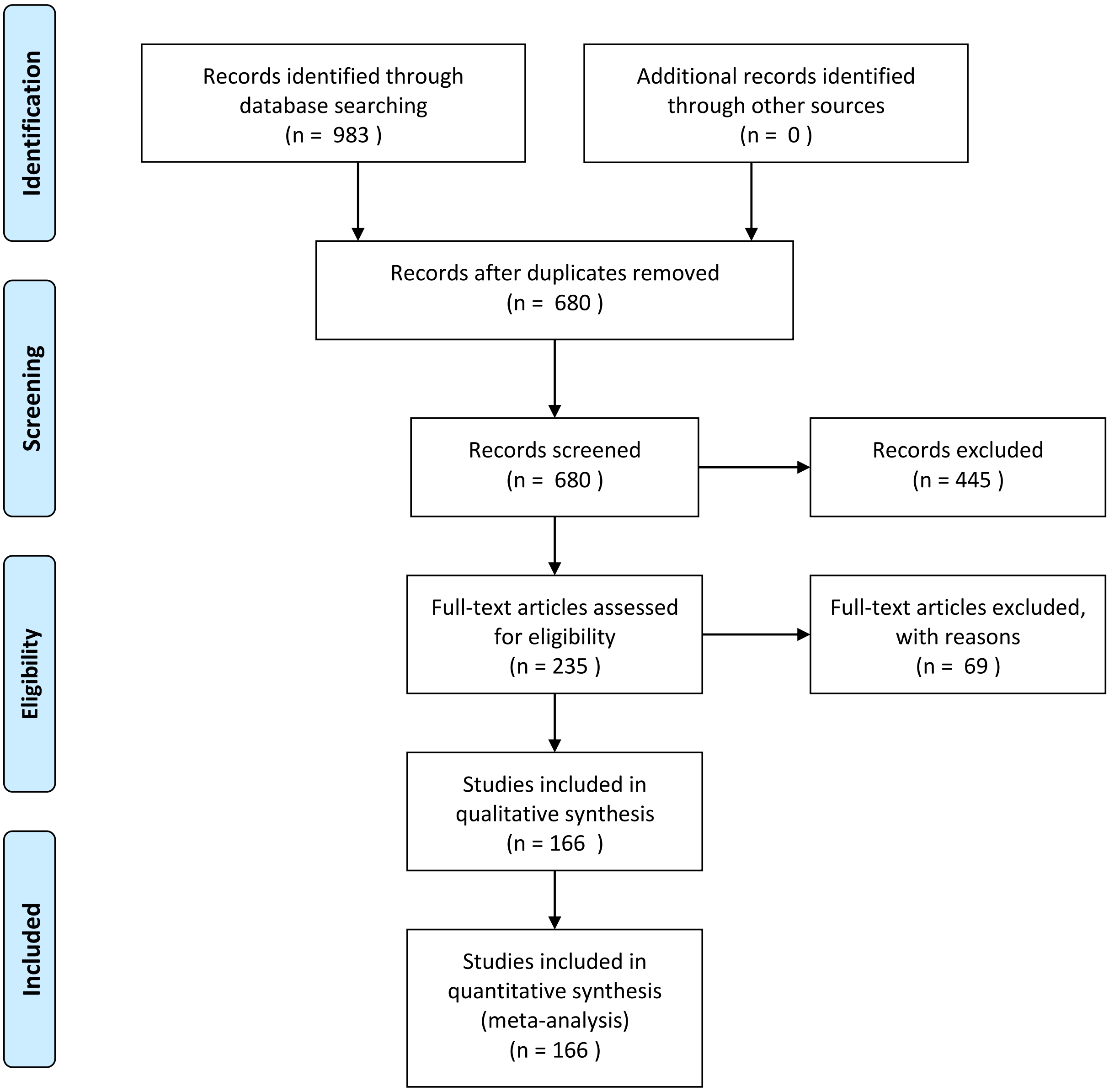

Fig. 2. Overview of the SLR process.

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

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	–

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.

4.5. 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 [24]. 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.

5. Results

The quantitative synthesis and analyses introduced in Section 4.5 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 [24].

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 3 shows the number of publications addressing personalization

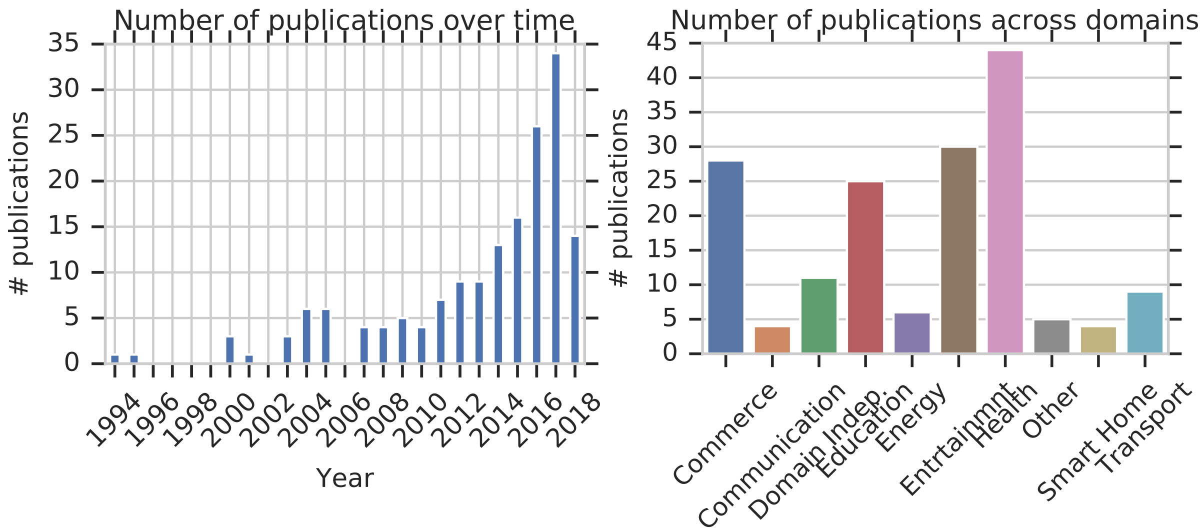

Fig. 3. Included papers over time and across domains. Note that only studies published prior to the query date of June 6, 2018 were included.

using RL techniques over time. A clear increase can be seen. With over forty entries, the health domain 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 4a 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 Figure 4b 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.

5.1. 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 trend and associated data collection. Figure 5a 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

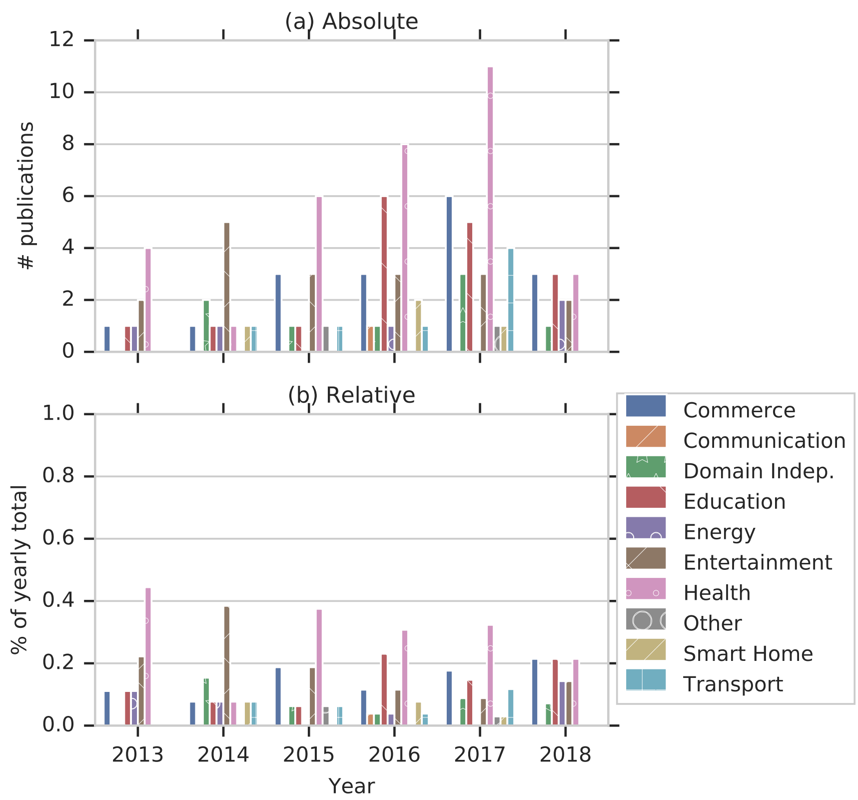

Fig. 4. Popularity of domains for the five most recent years.

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

hypothesize that in certain domains, such as health, safety is more frequently mentioned as a concern. Figure 5b 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 6 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.

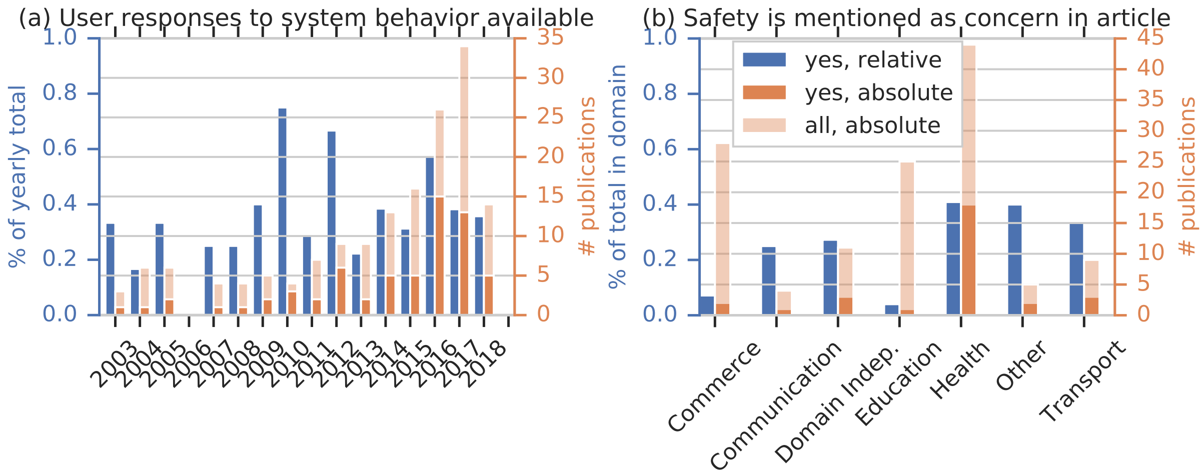

Fig. 5. Availability of user responses over time (a), and mentions of safety as a concern over domains (b).

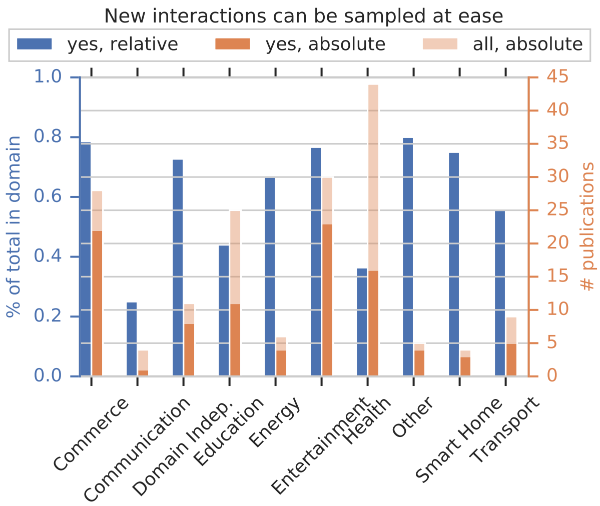

Fig. 6. New interactions with users can be sampled with ease.

Finally, we investigate whether upfront knowledge is available. In our analysis, we explore both real data as 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.

5.2. Solution

In our investigation into solutions, we first explore the algorithms that were used. Figure 7 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

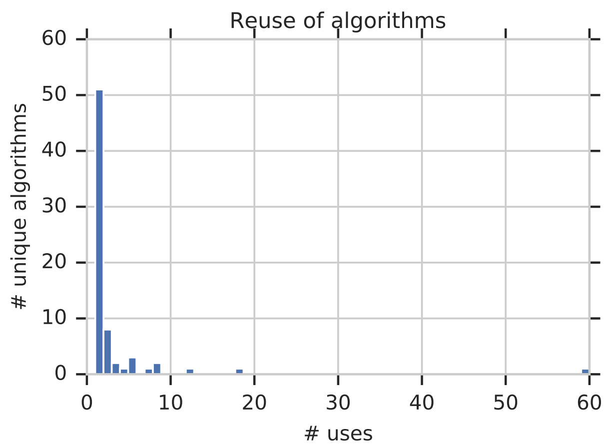

Fig. 7. Distribution of algorithm usage frequencies.

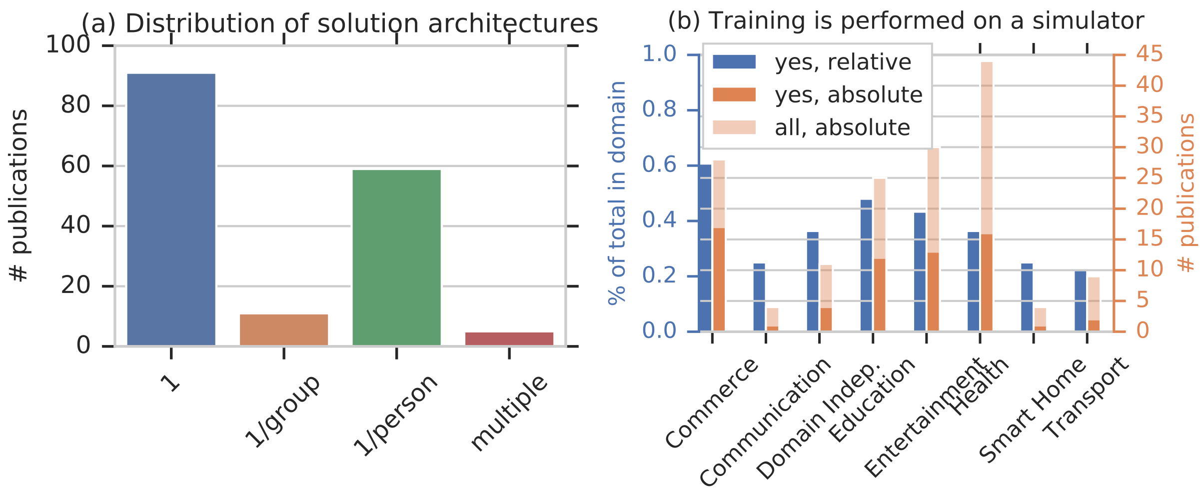

Fig. 8. Occurence of different solution architectures (a) and usage of simulators in training (b). For (a), publications that compare architectures are represented in the 'multiple' category.

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 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 8a 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,

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

Algorithm	# of uses
Q-learning	60
RL, not further specified	18
Contextual bandits	12
SARSA	8
Actor-critic	8
Inverse reinforcement learning	7
UCB	5
Policy iteration	5
LinUCB	5
Deep reinforcement learning	4
Fitted Q-iteration	3
DQN	3
Interactive reinforcement learning	2
TD-learning	2
DYNA-Q	2
Policy gradient	2
CLUB	2
Monte carlo	2
Thompson sampling	2
DDQN	2

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 8b 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 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.

5.3. Evaluation

In investigating evaluation rigor, we first turn to the data on which evaluations are based. Figure 9 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

Table 5

Number of models and the inclusion of user traits.

	Number of models			
Traits of users were used	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

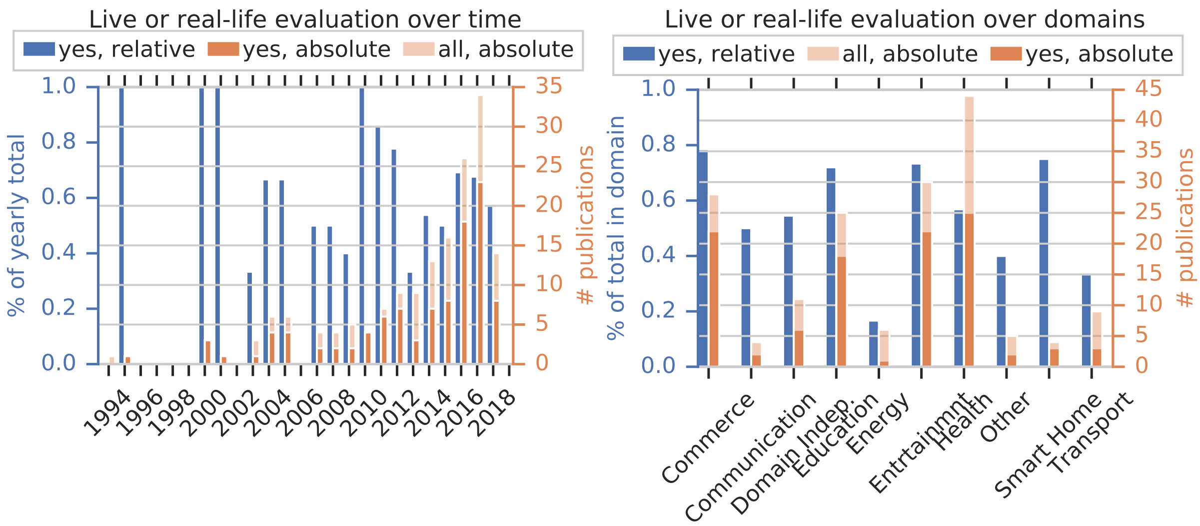

Fig. 9. Number of papers with a 'live' evaluation or evaluation using data on user responses to system behavior.

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 9 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 3. Indeed, there are 62 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 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.

Table 6
Comparison of settings with realistic and other evaluation.

	Evaluation	
	Realistic	Other
Data on user responses to system behavior are available	57 (.548)	9 (.145)
Safety is mentioned as a concern in the article	14 (.135)	16 (.258)
Models of user responses to system behavior are available	21 (.202)	20 (.323)
Privacy is mentioned as a concern in the article	7 (.067)	2 (.032)
New interactions with users can be sampled with ease	60 (.577)	37 (.597)
Total	104	63

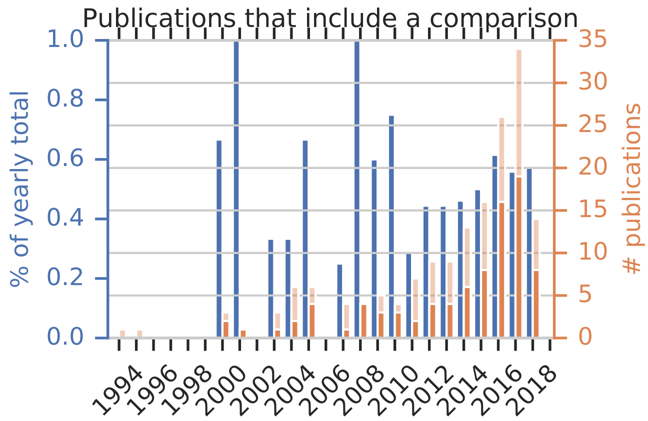

Fig. 10. Number of papers that include any comparison between solutions over time.

Figure 10 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.

6. 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 personalization 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 [53]. 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 [42, 54, 55]. Secondly, advances in the field of deep learning have wholly or partly removed the need for feature engineering [56]. 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 [39]. 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 [24].

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 [57, 58].

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 [20]. 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 [59].

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

Listing 1: Query for Scopus Database

```
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 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" 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 4: First Query for DBLP Database

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

Listing 5: Second Query for DBLP Database

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

Listing 6: First Query for Google Scholar 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 7: Second 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
```

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	[13, 18, 19, 53, 60–182]
	y	[183–221]
2	n	[62, 64, 69, 73, 78–80, 85, 87, 116, 119, 121, 131, 132, 137, 139, 170, 183, 184, 187, 188, 192, 194, 195, 198–200, 202, 203, 210, 214–218, 220]
	y	[13, 18, 19, 53, 60, 61, 63, 65–68, 70–72, 74–77, 81–84, 86, 88–115, 117, 118, 120, 122–130, 133–136, 138, 140–169, 171–182, 185, 186, 189–191, 193, 196, 197, 201, 204–209, 211–213, 219, 221]
3	n	[13, 18, 19, 60–67, 70–75, 77–80, 83, 86–90, 92–94, 96–101, 103, 105–107, 109, 111–115, 117–121, 123–126, 128–131, 133, 135–140, 142–158, 160–172, 174–192, 194, 195, 197–200, 202–206, 208–210, 215–221]
	y	[53, 68, 69, 76, 81, 82, 84, 85, 91, 95, 102, 104, 108, 110, 116, 122, 127, 132, 134, 141, 159, 173, 193, 196, 201, 207, 211–214]
4	n	[13, 18, 19, 53, 61–96, 98–104, 106–141, 143–146, 148–179, 181–194, 196–211, 213–215, 217–221]
	y	[60, 97, 105, 142, 147, 180, 195, 212, 216]
5	n	[13, 19, 53, 60, 62–64, 67, 69–75, 77, 79, 80, 83, 87–90, 93–101, 103–111, 113, 114, 117–119, 121–126, 129–134, 136–147, 149–151, 153–155, 157–160, 162–167, 170–172, 174–188, 193, 194, 196, 198–202, 204–207, 209, 210, 212–216, 218, 220]
	y	[18, 61, 65, 66, 68, 76, 78, 81, 82, 84–86, 91, 92, 102, 112, 115, 116, 120, 127, 128, 135, 148, 152, 156, 161, 168, 169, 173, 189–192, 195, 197, 203, 208, 211, 217, 219, 221]
6	n	[18, 60–62, 64–66, 69, 71–73, 75, 76, 79, 80, 82–89, 91–96, 98–104, 110–116, 118, 120–122, 125–131, 134, 135, 137, 139–142, 144–146, 148, 149, 151, 158–161, 164–166, 171, 173–178, 182, 183, 185, 187, 188, 192, 193, 196, 197, 199, 201, 203, 205, 208, 210, 211, 213, 215, 221]
	y	[13, 19, 53, 63, 67, 68, 70, 74, 77, 78, 81, 90, 97, 105–109, 117, 119, 123, 124, 132, 133, 136, 138, 143, 147, 150, 152–157, 162, 163, 167–170, 172, 179–181, 184, 186, 189–191, 194, 195, 198, 200, 202, 204, 206, 207, 209, 212, 214, 216–220]
7	n	[19, 60, 63–65, 72–81, 86–88, 90, 91, 93, 100, 102–104, 108, 110, 113–116, 121, 122, 124, 126, 128, 131, 133–135, 137, 139, 142, 143, 145, 146, 158–160, 164, 171, 173, 175–177, 187, 188, 196, 198, 199, 202, 203, 207, 209, 210, 212–214, 218]
	y	[13, 18, 53, 61, 62, 66–71, 82–85, 89, 92, 94–99, 101, 105–107, 109, 111, 112, 117–120, 123, 125, 127, 129, 130, 132, 136, 138, 140, 141, 144, 147–157, 161–163, 165–170, 172, 174, 178–186, 189–195, 197, 200, 201, 204–206, 208, 211, 215–217, 219–221]
8	n	[68, 70, 73, 75, 78, 89, 91, 97, 105, 109, 113, 123, 129, 132, 137, 140, 150, 153, 157, 162–164, 166, 172, 178, 180, 186, 193, 194, 200, 201, 213, 216, 219]

#	Value	Publications
	y	[13, 18, 19, 53, 60–67, 69, 71, 72, 74, 76, 77, 79–88, 90, 92–96, 98–104, 106–108, 110–112, 114–122, 124–128, 130, 131, 133–136, 138, 139, 141–149, 151, 152, 154–156, 158–161, 165, 167–171, 173–177, 179, 181–185, 187–192, 195–199, 202–212, 214, 215, 217, 218, 220, 221]
10	1	[53, 60–63, 66, 71–75, 77, 79–82, 85, 86, 88–91, 93–97, 99, 100, 102, 103, 105, 107, 108, 113, 115–117, 122, 124, 127–129, 131, 135–138, 141–143, 145, 146, 148–155, 157, 159–163, 167–169, 171, 173, 175, 176, 178, 181, 182, 186, 197–200, 208, 210–212, 214–218]
	1/group	[18, 65, 83, 106, 109, 111, 114, 133, 164, 165, 177]
	1/person	[13, 19, 64, 67–70, 76, 78, 84, 92, 98, 101, 104, 110, 112, 118–121, 125, 126, 130, 132, 134, 139, 140, 144, 156, 158, 166, 170, 172, 174, 179, 184, 185, 187–196, 201–207, 209, 213, 219–221]
	multiple	[87, 123, 147, 180, 183]
11	not used	[53, 62, 64–66, 69, 71, 73, 75, 77, 79–81, 83–85, 88, 90, 92–95, 97, 99–101, 103, 109, 113, 117, 118, 122, 124, 125, 129–131, 137, 141, 143, 145, 146, 148, 150, 151, 156, 158, 161, 166, 167, 174, 179, 181, 184–186, 188, 192, 194, 197–201, 204, 207, 210, 214–218, 221]
	other	[19, 67, 70, 120, 121, 123, 134, 136, 153, 155, 171, 173, 180, 183, 195, 203, 206]
	state representation	[13, 18, 60, 61, 63, 68, 72, 74, 76, 78, 82, 86, 87, 89, 91, 96, 98, 102, 104–108, 110–112, 114–116, 119, 126–128, 132, 133, 135, 138–140, 142, 144, 147, 149, 152, 154, 157, 159, 160, 162–165, 168–170, 172, 175–178, 182, 187, 189–191, 193, 196, 202, 205, 208, 209, 211–213, 219, 220]
12	batch	[18, 53, 60, 77, 80, 87, 89, 97, 102, 106, 110, 113–116, 119, 127, 128, 135, 136, 139, 147–151, 154–157, 159, 162, 163, 167, 171, 175–178, 180, 182, 186, 187, 189–192, 195, 198, 203, 209, 219]
	n	[166]
	online	[13, 66–71, 74, 76, 78, 82, 85, 86, 91–93, 96, 98–101, 104, 105, 107–109, 112, 118, 120–122, 124, 125, 129, 130, 132, 140–144, 146, 152, 153, 161, 168–170, 174, 179, 181, 184, 185, 188, 193, 194, 196, 199–202, 204–208, 211, 213, 214, 216–218, 220, 221]
	other	[19, 84, 94, 123, 126, 133, 164, 172]
	unknown	[61–65, 72, 73, 75, 79, 81, 83, 88, 90, 95, 103, 111, 117, 131, 134, 137, 138, 145, 158, 160, 165, 173, 183, 197, 210, 212, 215]
13	n	[13, 19, 53, 60, 63, 67, 71–75, 77, 78, 80, 81, 87, 88, 90, 97, 99, 100, 102, 103, 105–108, 111, 113–119, 122, 124, 127–131, 133, 135, 136, 138, 139, 142–147, 154–157, 160, 163, 165–169, 171–173, 175, 178, 179, 181, 183–186, 198–200, 202, 204, 206, 209, 212, 214, 215, 217, 218, 220]
	y	[18, 61, 62, 64–66, 68–70, 76, 79, 82–86, 89, 91–96, 98, 101, 104, 109, 110, 112, 120, 121, 123, 125, 126, 132, 134, 137, 140, 141, 148–153, 158, 159, 161, 162, 164, 170, 174, 176, 177, 180, 182, 187–197, 201, 203, 205, 207, 208, 210, 211, 213, 216, 219, 221]
14	n	[13, 18, 19, 61, 62, 64–66, 68–76, 78, 79, 82–85, 89–93, 95, 96, 98–101, 103, 104, 107, 109–112, 120–122, 125, 126, 129, 131, 132, 134, 137, 140–143, 146, 150, 158, 160–162, 164, 166, 170, 174, 176–179, 181, 183–186, 188–195, 197, 201, 203–208, 210–215, 217–219, 221]

#	Value	Publications
	y	[53, 60, 63, 67, 77, 80, 81, 86–88, 94, 97, 102, 105, 106, 108, 113–119, 123, 124, 127, 128, 130, 133, 135, 136, 138, 139, 144, 145, 147–149, 151–157, 159, 163, 165, 167–169, 171–173, 175, 180, 182, 187, 196, 198–200, 202, 209, 216, 220]
15	n	[18, 53, 60, 62–67, 69, 70, 72, 73, 75–77, 79–89, 91–93, 95, 97, 98, 101, 102, 105, 106, 108–121, 123–128, 130–137, 139–141, 144, 145, 147–153, 155, 157–169, 171, 173, 175–178, 180, 182, 183, 186–188, 192, 193, 195–200, 202, 203, 205, 207–209, 211–213, 215, 219–221]
	y	[13, 19, 61, 68, 71, 74, 78, 90, 94, 96, 99, 100, 103, 104, 107, 122, 129, 138, 142, 143, 146, 154, 156, 170, 172, 174, 179, 181, 184, 185, 189–191, 194, 201, 204, 206, 210, 214, 216–218]
16	n	[13, 19, 53, 60, 63, 67, 71–75, 77, 78, 80, 87, 88, 90, 96, 97, 99, 100, 102, 103, 105–108, 111, 113–119, 122, 124, 127–131, 133, 135, 136, 138, 139, 142–147, 154–157, 160, 163, 165–169, 172, 173, 175, 178, 179, 181, 183–186, 196, 198, 200, 202, 204, 206, 209, 212–215, 217, 218, 220]
	y	[18, 61, 62, 64–66, 68–70, 76, 79, 81–86, 89, 91–95, 98, 101, 104, 109, 110, 112, 120, 121, 123, 125, 126, 132, 134, 137, 140, 141, 148–153, 158, 159, 161, 162, 164, 170, 171, 174, 176, 177, 180, 182, 187–195, 197, 199, 201, 203, 205, 207, 208, 210, 211, 216, 219, 221]
17	n	[13, 18, 19, 61, 62, 64–66, 68–79, 81–85, 89–93, 95, 96, 98–101, 103, 104, 107, 109–112, 120–122, 124–126, 129, 131, 132, 134, 137, 140–143, 146, 147, 149, 150, 156, 158, 160–162, 164, 166, 170, 171, 174, 176, 178, 179, 181, 183–186, 188–195, 197, 199, 201, 203–208, 210–215, 217–219, 221]
	y	[53, 60, 63, 67, 80, 86–88, 94, 97, 102, 105, 106, 108, 113–119, 123, 127, 128, 130, 133, 135, 136, 138, 139, 144, 145, 148, 151–155, 157, 159, 163, 165, 167–169, 172, 173, 175, 177, 180, 182, 187, 196, 198, 200, 202, 209, 216, 220]
18	n	[18, 53, 60, 62–67, 69, 70, 72, 73, 75, 76, 79–89, 91–93, 95, 97, 98, 101, 102, 105, 106, 108–121, 123, 125–128, 130–137, 139–141, 144, 145, 149–153, 155, 157–167, 171, 173, 175–177, 182, 183, 187, 188, 192, 193, 195, 196, 198, 200, 202, 203, 205, 208, 209, 211–213, 215, 219–221]
	y	[13, 19, 61, 68, 71, 74, 77, 78, 90, 94, 96, 99, 100, 103, 104, 107, 122, 124, 129, 138, 142, 143, 146–148, 154, 156, 168–170, 172, 174, 178–181, 184–186, 189–191, 194, 197, 199, 201, 204, 206, 207, 210, 214, 216–218]
19	n	[53, 60, 66, 67, 69–75, 77, 78, 80–83, 85, 88, 89, 91, 94–100, 103, 106–114, 117–122, 124, 125, 128, 129, 131–134, 136, 139–143, 145, 146, 148–152, 157, 158, 160–162, 164–167, 174–176, 178–180, 182, 183, 187–191, 193, 195, 196, 198, 200–203, 205–208, 211–213, 215, 217–219, 221]
	y	[13, 18, 19, 61–65, 68, 76, 79, 84, 86, 87, 90, 92, 93, 101, 102, 104, 105, 115, 116, 123, 126, 127, 130, 135, 137, 138, 144, 147, 153–156, 159, 163, 168–173, 177, 181, 184–186, 192, 194, 197, 199, 204, 209, 210, 214, 216, 220]
20	n	[19, 53, 60–64, 66, 67, 69–76, 78–85, 87–90, 92–99, 102, 103, 105–108, 110–118, 121, 122, 124–127, 129, 131–134, 137, 139–141, 145–148, 152, 153, 155–158, 160, 162–167, 171, 174, 176, 177, 179–181, 183–196, 201, 204, 206–208, 210–215, 217–219, 221]

#	Value	Publications
	y	[13, 18, 65, 68, 77, 86, 91, 100, 101, 104, 109, 119, 120, 123, 128, 130, 135, 136, 138, 142–144, 149–151, 154, 159, 161, 168–170, 172, 173, 175, 178, 182, 197–200, 202, 203, 205, 209, 216, 220]
Domain	Commerce	[53, 60, 67, 72, 87, 89, 94, 96, 106, 111, 112, 119, 123, 129, 138, 151, 154, 155, 157, 164, 170, 171, 175, 178, 200, 201, 217, 220]
	Communication	[81, 101, 103, 142]
	Domain Independent	[68, 70, 78, 116, 125, 126, 132, 152, 168, 169, 219]
	Education	[18, 19, 74, 75, 77, 88, 90, 99, 100, 109, 113, 114, 131, 139, 145, 148, 149, 162, 163, 167, 196–198, 206, 208]
	Energy	[98, 120, 121, 161, 174, 203]
	Entertainment	[61, 62, 66, 79, 93, 97, 105, 117, 118, 124, 136, 137, 140, 144, 147, 150, 153, 172, 180, 186, 189–192, 194, 199, 204, 210, 216, 218]
	Health	[63, 73, 80, 82–86, 91, 92, 102, 104, 108, 110, 115, 122, 127, 128, 133–135, 141, 143, 158–160, 166, 173, 176, 177, 179, 181–184, 187, 195, 202, 207, 209, 211, 212, 214, 221]
	Other	[13, 69, 71, 213, 215]
	Smart Home	[107, 156, 185, 188]
	Transport	[64, 65, 76, 95, 130, 146, 165, 193, 205]

References

[1] G.S. Ginsburg and J.J. McCarthy, Personalized medicine: revolutionizing drug discovery and patient care, *TRENDS in Biotechnology* **19**(12) (2001), 491–496.

[2] M.G. Aspinall and R.G. Hamermesh, Realizing the promise of personalized medicine, *Harvard business review* **85**(10) (2007), 108.

[3] P. Maes and R. Kozierok, Learning interface agents, in: *AAAI*, Vol. 93, 1993, pp. 459–465.

[4] 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.

[5] P. Resnick and H.R. Varian, Recommender systems, *Communications of the ACM* **40**(3) (1997), 56–58.

[6] F. Ricci, L. Rokach and B. Shapira, Introduction to recommender systems handbook, in: *Recommender systems handbook*, Springer, 2011, pp. 14–17.

[7] 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.

[8] B.J. Pine, B. Victor and A.C. Boynton, Making mass customization work, *Harvard business review* **71**(5) (1993), 108–11.

[9] 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.

[10] 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.

[11] A. Barto, P. Thomas and R. Sutton, Some Recent Applications of Reinforcement Learning (2017).

[12] F.M. Harper, X. Li, Y. Chen and J.A. Konstan, An economic model of user rating in an online recommender system, *Lecture notes in computer science* **3538** (2005), 307.

[13] 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.

[14] 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.

[15] B.-T. Zhang and Y.-W. Seo, Personalized web-document filtering using reinforcement learning, *Applied Artificial Intelligence* **15**(7) (2001), 665–685.

[16] 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.

[17] Y. Zhao, M.R. Kosorok and D. Zeng, Reinforcement learning design for cancer clinical trials, *Statistics in medicine* **28**(26) (2009), 3294–3315.

[18] 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.

[19] 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.

[20] J. Kober and J. Peters, Reinforcement Learning in Robotics: A Survey, in: *Reinforcement learning*, Vol. 12, M. Wiering and M. Van Otterlo, eds, Springer, 2012, pp. "596–597".

[21] 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.

[22] 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.

[23] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang and W. Zaremba, Openai gym, *arXiv preprint arXiv:1606.01540* (2016).

[24] 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.

[25] R.S. Sutton and A.G. Barto, *Reinforcement learning: An introduction*, MIT press Cambridge, 1998.

[26] M. Wiering and M. Van Otterlo, Reinforcement learning, *Adaptation, learning, and optimization* **12** (2012).

[27] C. Szepesvári, Algorithms for reinforcement learning, *Synthesis lectures on artificial intelligence and machine learning* **4**(1) (2010), 1–103.

[28] 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.

[29] R.E. Bellman, *Adaptive control processes: a guided tour*, Vol. 2045, Princeton university press, 2015.

[30] 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.

[31] 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.

[32] 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.

[33] S.J. Pan and Q. Yang, A survey on transfer learning, *IEEE Transactions on knowledge and data engineering* **22**(10) (2010), 1345–1359.

[34] D. Riecken, Personalized views of personalization, *Communications of the ACM* **43**(8) (2000), 26–26.

[35] 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.

[36] J.B. Schafer, D. Frankowski, J. Herlocker and S. Sen, Collaborative filtering recommender systems, in: *The adaptive web*, Springer, 2007, pp. 291–324.

[37] 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*, ACM, 2010, pp. 47–51.

[38] 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.

[39] J. Garcıa and F. Fernández, A comprehensive survey on safe reinforcement learning, *Journal of Machine Learning Research* **16**(1) (2015), 1437–1480.

[40] T. Hester and P. Stone, Learning and Using Models, in: *Reinforcement learning*, Vol. 12, M. Wiering and M. Van Otterlo, eds, Springer, 2012, p. "120".

[41] L.-J. Lin, Self-improving reactive agents based on reinforcement learning, planning and teaching, *Machine learning* **8**(3–4) (1992), 293–321.

[42] P.S. Thomas, G. Theocharous and M. Ghavamzadeh, High-confidence off-policy evaluation, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[43] 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* (2006), 13–28.

[44] P. Brusilovski, A. Kobsa and W. Nejdl, *The adaptive web: methods and strategies of web personalization*, Vol. 4321, Springer Science & Business Media, 2007.

[45] M.K. Khribi, M. Jemni and O. Nasraoui, Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval, in: *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on*, IEEE, 2008, pp. 241–245.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] N. Jiang and L. Li, Doubly Robust Off-policy Value Evaluation for Reinforcement Learning, in: *International Conference on Machine Learning*, 2016, pp. 652–661.

[56] 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.

[57] 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.

[58] E. Ie, C.-w. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu and C. Boutilier, RecSim: A Configurable Simulation Platform for Recommender Systems, *arXiv preprint arXiv:1909.04847* (2019).

[59] S. Russell, P. Norvig and A. Intelligence, A modern approach, *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* **25** (1995), 4.

[60] N. Abe, N. Verma, C. Apte and R. Schroko, Cross channel optimized marketing by reinforcement learning, *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04* (2004).

[61] G. Andrade, G. Ramalho, H. Santana and V. Corruble, Automatic computer game balancing, *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems - AAMAS '05* (2005).

[62] N. Bagdure and B. Ambudkar, Reducing Delay during Vertical Handover, *2015 International Conference on Computing Communication Control and Automation* (2015).

[63] 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.

[64] A.L.C. Bazzan, Synergies between evolutionary computation and multiagent reinforcement learning, *Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO '17* (2017).

[65] H. Bi, O.J. Akinwande and E. Gelenbe, Emergency Navigation in Confined Spaces Using Dynamic Grouping, *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies* (2015).

[66] A. Bodas, B. Upadhyay, C. Nadiger and S. Abdelhak, Reinforcement learning for game personalization on edge devices, *2018 International Conference on Information and Computer Technologies (ICICT)* (2018).

[67] D. Bouneffouf, A. Bouzeghoub and A.L. Gançarski, Hybrid-ϵ-greedy for Mobile Context-Aware Recommender System, *Lecture Notes in Computer Science* (2012), 468–479.

[68] J. Bragg, Mausam and D.S. Weld, Optimal Testing for Crowd Workers, in: *AAMAS*, 2016.

[69] A.B. Buduru and S.S. Yau, An Effective Approach to Continuous User Authentication for Touch Screen Smart Devices, *2015 IEEE International Conference on Software Quality, Reliability and Security* (2015).

[70] 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.

[71] 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, *2011 IEEE International Conference on Robotics and Biomimetics* (2011).

[72] L. Cella, Modelling User Behaviors with Evolving Users and Catalogs of Evolving Items, *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17* (2017).

[73] B. Chakraborty and S.A. Murphy, Dynamic Treatment Regimes, *Annual Review of Statistics and Its Application* **1**(1) (2014), 447–464.

[74] J. Chan and G. Nejat, A learning-based control architecture for an assistive robot providing social engagement during cognitively stimulating activities, *2011 IEEE International Conference on Robotics and Automation* (2011).

[75] J. Chen and Z. Yang, A learning multi-agent system for personalized information fiftering, *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint* (2003).

[76] X. Chen, Y. Zhai, C. Lu, J. Gong and G. Wang, A learning model for personalized adaptive cruise control, *2017 IEEE Intelligent Vehicles Symposium (IV)* (2017).

[77] M. Chi, K. VanLehn, D. Litman and P. Jordan, Inducing Effective Pedagogical Strategies Using Learning Context Features, *Lecture Notes in Computer Science* (2010), 147–158.

[78] 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, *2014 IEEE International Workshop on Advanced Robotics and its Social Impacts* (2014).

[79] 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.

[80] M. Daltayanni, C. Wang and R. Akella, A Fast Interactive Search System for Healthcare Services, *2012 Annual SRII Global Conference* (2012).

[81] 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.
[82] E. Daskalaki, P. Diem and S.G. Mougiakakou, Personalized tuning of a reinforcement learning control algorithm for glucose regulation, *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2013).
[83] 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.
[84] 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.
[85] 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.
[86] K. Deng, J. Pineau and S. Murphy, Active learning for personalizing treatment, *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)* (2011).
[87] A.A. Deshmukh, Ürün Dogan and C. Scott, Multi-Task Learning for Contextual Bandits, in: *NIPS*, 2017.
[88] M. El Fouki, N. Aknin and K.E. El. Kadiri, Intelligent Adapted e-Learning System based on Deep Reinforcement Learning, *Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems - ICCWCS'17* (2017).
[89] J. Feng, H. Li, M. Huang, S. Liu, W. Ou, Z. Wang and X. Zhu, Learning to Collaborate, *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (2018).
[90] 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.
[91] 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, *Fourth International Conference on Machine Learning and Applications (ICMLA'05)* (2005).
[92] A.E. Gaweda, Improving management of Anemia in End Stage Renal Disease using Reinforcement Learning, *2009 International Joint Conference on Neural Networks* (2009).
[93] B.S. Ghahfarokhi and N. Movahhedinia, A personalized QoE-aware handover decision based on distributed reinforcement learning, *Wireless Networks* **19**(8) (2013), 1807–1828.
[94] D.N. Hill, H. Nassif, Y. Liu, A. Iyer and S.V.N. Vishwanathan, An Efficient Bandit Algorithm for Realtime Multivariate Optimization, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* (2017).
[95] Z. Huajun, Z. Jin, W. Rui and M. Tan, Multi-objective reinforcement learning algorithm and its application in drive system, *2008 34th Annual Conference of IEEE Industrial Electronics* (2008).
[96] S.-l. Huang and F.-r. Lin, Designing intelligent sales-agent for online selling, *Proceedings of the 7th international conference on Electronic commerce - ICEC '05* (2005).
[97] S. Jaradat, N. Dokoohaki, M. Matskin and E. Ferrari, Trust and privacy correlations in social networks: A deep learning framework, *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2016).
[98] Z. Jin and Z. Huajun, Multi-objective reinforcement learning algorithm and its improved convergency method, *2011 6th IEEE Conference on Industrial Electronics and Applications* (2011).
[99] A.A. Kardan and O.R.B. Speily, Smart Lifelong Learning System Based on Q-Learning, *2010 Seventh International Conference on Information Technology: New Generations* (2010).
[100] I. Kastanis and M. Slater, Reinforcement learning utilizes proxemics, *ACM Transactions on Applied Perception* **9**(1) (2012), 1–15.
[101] I. Koukoutsidis, A learning strategy for paging in mobile environments, *5th European Personal Mobile Communications Conference 2003* (2003).
[102] E.F. Krakow, M. Hemmer, T. Wang, B. Logan, M. Arora, S. Spellman, D. Couriel, A. Alousi, J. Pidala, M. Last and 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.
[103] G. Lee, S. Bauer, P. Faratin and J. Wroclawski, Learning user preferences for wireless services provisioning, *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.* (2004), 480–487.
[104] K. Li and M.Q.-H. Meng, Personalizing a Service Robot by Learning Human Habits from Behavioral Footprints, *Engineering* **1**(1) (2015), 079–084.
[105] L. Li, W. Chu, J. Langford and R.E. Schapire, A contextual-bandit approach to personalized news article recommendation, *Proceedings of the 19th international conference on World wide web - WWW '10* (2010).

[106] Z. Li, J. Kiseleva, M. de Rijke and A. Grotov, Towards Learning Reward Functions from User Interactions, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR '17* (2017).
[107] J. Lim, H. Son, D. Lee and D. Lee, An MARL-Based Distributed Learning Scheme for Capturing User Preferences in a Smart Environment, *2017 IEEE International Conference on Services Computing (SCC)* (2017).
[108] Y. Liu, B. Logan, N. Liu, Z. Xu, J. Tang and Y. Wang, Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data, *2017 IEEE International Conference on Healthcare Informatics (ICHI)* (2017).
[109] 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.
[110] C. Lowery and A.A. Faisal, Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control, *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* (2013).
[111] O. Madani and D. DeCoste, Contextual recommender problems [extended abstract], *Proceedings of the 1st international workshop on Utility-based data mining - UBDM '05* (2005).
[112] T. Mahmood and F. Ricci, Learning and adaptivity in interactive recommender systems, *Proceedings of the ninth international conference on Electronic commerce - ICEC '07* (2007).
[113] A. Malpani, B. Ravindran and H. Murthy, Personalized Intelligent Tutoring System Using Reinforcement Learning, in: *FLAIRS Conference*, 2011.
[114] I. Manickam, A.S. Lan and R.G. Baraniuk, Contextual multi-armed bandit algorithms for personalized learning action selection, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017).
[115] 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.
[116] 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, *Recent Advances in Reinforcement Learning* (2008), 191–204.
[117] D. Massimo, M. Elahi and F. Ricci, Learning User Preferences by Observing User-Items Interactions in an IoT Augmented Space, *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17* (2017).
[118] K. Masumitsu and T. Echigo, Video summarization using reinforcement learning in eigenspace, *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)* (2000).
[119] B.C. May, N. Korda, A. Lee and D.S. Leslie, Optimistic Bayesian sampling in contextual-bandit problems, *Journal of Machine Learning Research* **13**(Jun) (2012), 2069–2106.
[120] E. Mengelkamp, J. Gärttner and C. Weinhardt, Intelligent Agent Strategies for Residential Customers in Local Electricity Markets, *Proceedings of the Ninth International Conference on Future Energy Systems - e-Energy '18* (2018).
[121] E. Mengelkamp and C. Weinhardt, Clustering Household Preferences in Local Electricity Markets, *Proceedings of the Ninth International Conference on Future Energy Systems - e-Energy '18* (2018).
[122] N. Merkle and S. Zander, Agent-Based Assistance in Ambient Assisted Living Through Reinforcement Learning and Semantic Technologies, *Lecture Notes in Computer Science* (2017), 180–188.
[123] K. Mo, Y. Zhang, S. Li, J. Li and Q. Yang, Personalizing a Dialogue System With Transfer Reinforcement Learning, in: *AAAI*, 2018.
[124] O. Moling, L. Baltrunas and F. Ricci, Optimal radio channel recommendations with explicit and implicit feedback, *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12* (2012).
[125] A. Moon, T. Kang, H. Kim and H. Kim, A Service Recommendation Using Reinforcement Learning for Network-based Robots in Ubiquitous Computing Environments, *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication* (2007).
[126] S. Narvekar, J. Sinapov and P. Stone, Autonomous Task Sequencing for Customized Curriculum Design in Reinforcement Learning, in: *IJCAI*, 2017.
[127] S. Nemati, M.M. Ghassemi and G.D. Clifford, Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach, *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2016).
[128] D. Neumann, T. Mansi, L. Itu, B. Georgescu, E. Kayvanpour, F. Sedaghat-Hamedani, A. Amr, J. Haas, H. Katus, B. Meder and et al., A self-taught artificial agent for multi-physics computational model personalization, *Medical Image Analysis* **34** (2016), 52–64.
[129] D. Oh and C.L. Tan, Making Better Recommendations with Online Profiling Agents, *AI Magazine* **26** (2004), 29–40.
[130] 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.
[131] V. Pant, S. Bhasin and S. Jain, Self-learning system for personalized e-learning, *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)* (2017).

[132] P. Patompak, S. Jeong, I. Nilkhamhang and N.Y. Chong, Learning social relations for culture aware interaction, *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)* (2017).
[133] 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.
[134] 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.
[135] 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, *CoRR* **abs/1704.06300** (2017).
[136] M. Preda and D. Popescu, Personalized Web Recommendations: Supporting Epistemic Information about End-Users, *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)* (2005).
[137] 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.
[138] Z. Qin, I. Rishabh and J. Carnahan, A Scalable Approach for Periodical Personalized Recommendations, *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16* (2016).
[139] E. Rennison, Personalized Galaxies of Infomation, in: *Companion of the ACM Conference on Human Factors in Computing Systems (CHI'95)*, 1995.
[140] H. Ritschel and E. André, Real-Time Robot Personality Adaptation based on Reinforcement Learning and Social Signals, *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (2017).
[141] M. Rudary, S. Singh and M.E. Pollack, Adaptive cognitive orthotics, *Twenty-first international conference on Machine learning - ICML '04* (2004).
[142] S. Saha and R. Quazi, Emotion-driven learning agent for setting rich presence in mobile telephony, *2008 11th International Conference on Computer and Information Technology* (2008).
[143] Y.A. Sekhavat, MPRL: Multiple-Periodic Reinforcement Learning for difficulty adjustment in rehabilitation games, *2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH)* (2017).
[144] 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.
[145] S. Shen and M. Chi, Reinforcement Learning, *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16* (2016).
[146] A.R. Srinivasan and S. Chakraborty, Path planning with user route preference - A reward surface approximation approach using orthogonal Legendre polynomials, *2016 IEEE International Conference on Automation Science and Engineering (CASE)* (2016).
[147] A. Srivihok and P. Sukonmanee, Intelligent Agent for e-Tourism: Personalization Travel Support Agent using Reinforcement Learning, in: *WWW 2005*, 2005.
[148] P.-H. Su, C.-H. Wu and L.-S. Lee, A Recursive Dialogue Game for Personalized Computer-Aided Pronunciation Training, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2014), 1–1.
[149] P.-h. Su, Y.-B. Wang, T.-h. Yu and L.-s. Lee, A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013).
[150] N. Taghipour and A. Kardan, A hybrid web recommender system based on Q-learning, *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08* (2008).
[151] N. Taghipour, A. Kardan and S.S. Ghidary, Usage-based web recommendations, *Proceedings of the 2007 ACM conference on Recommender systems - RecSys '07* (2007).
[152] L. Tang, Y. Jiang, L. Li and T. Li, Ensemble contextual bandits for personalized recommendation, *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14* (2014).
[153] L. Tang, Y. Jiang, L. Li, C. Zeng and T. Li, Personalized Recommendation via Parameter-Free Contextual Bandits, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15* (2015).
[154] L. Tang, R. Rosales, A. Singh and D. Agarwal, Automatic ad format selection via contextual bandits, *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13* (2013).
[155] M. Tavakol and U. Brefeld, A Unified Contextual Bandit Framework for Long- and Short-Term Recommendations, *Lecture Notes in Computer Science* (2017), 269–284.
[156] B. Tegelund, H. Son and D. Lee, A task-oriented service personalization scheme for smart environments using reinforcement learning, *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)* (2016).
[157] G. Theocharous, P.S. Thomas and M. Ghavamzadeh, Ad Recommendation Systems for Life-Time Value Optimization, *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion* (2015).

[158] S. Triki and C. Hanachi, A Self-adaptive System for Improving Autonomy and Public Spaces Accessibility for Elderly, *Smart Innovation, Systems and Technologies* (2017), 53–66.
[159] 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.
[160] K. Tsiakas, C. Abellanoza and F. Makedon, Interactive Learning and Adaptation for Robot Assisted Therapy for People with Dementia, *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '16* (2016).
[161] D. Urieli and P. Stone, TacTex'13: a champion adaptive power trading agent, in: *AAMAS*, 2014.
[162] P. Wang, J.P. Rowe, W. Min, B.W. Mott and J.C. Lester, Interactive Narrative Personalization with Deep Reinforcement Learning, in: *IJCAI*, 2017.
[163] P. Wang, J. Rowe, B. Mott and J. Lester, Decomposing Drama Management in Educational Interactive Narrative: A Modular Reinforcement Learning Approach, *Lecture Notes in Computer Science* (2016), 270–282.
[164] X. Wang, M. Zhang, F. Ren and T. Ito, GongBroker: A Broker Model for Power Trading in Smart Grid Markets, *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (2015).
[165] G. Wu, Y. Ding, Y. Li, J. Luo, F. Zhang and J. Fu, Data-driven inverse learning of passenger preferences in urban public transits, *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* (2017).
[166] J. Xu, T. Xing and M. van der Schaar, Personalized Course Sequence Recommendations, *IEEE Transactions on Signal Processing* **64**(20) (2016), 5340–5352.
[167] 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.
[168] 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.
[169] M. Yang, Z. Zhao, W. Zhao, X. Chen, J. Zhu, L. Zhou and Z. Cao, Personalized Response Generation via Domain adaptation, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17* (2017).
[170] 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.
[171] C. Zeng, Q. Wang, S. Mokhtari and T. Li, Online Context-Aware Recommendation with Time Varying Multi-Armed Bandit, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016).
[172] B.-T. Zhang and Y.-W. Seo, Personalized web-document filtering using reinforcement learning, *Applied Artificial Intelligence* **15**(7) (2001), 665–685.
[173] Y. Zhang, R. Chen, J. Tang, W.F. Stewart and J. Sun, LEAP, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* (2017).
[174] 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, *2013 IEEE International Conference on Automation Science and Engineering (CASE)* (2013).
[175] Y. Zhao, S. Wang, Y. Zou, J. Ng and T. Ng, Automatically Learning User Preferences for Personalized Service Composition, *2017 IEEE International Conference on Web Services (ICWS)* (2017).
[176] Y. Zhao, M.R. Kosorok and D. Zeng, Reinforcement learning design for cancer clinical trials, *Statistics in Medicine* **28**(26) (2009), 3294–3315.
[177] 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.
[178] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N.J. Yuan, X. Xie and Z. Li, DRN, *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (2018).
[179] H. Zheng and J. Jumadinova, OWLS: Observational Wireless Life-enhancing System (Extended Abstract), in: *AAMAS*, 2016.
[180] L. Zhou and E. Brunskill, Latent Contextual Bandits and their Application to Personalized Recommendations for New Users, in: *IJCAI*, 2016.
[181] 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.
[182] 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.
[183] 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.
[184] A. Atrash and J. Pineau, A bayesian reinforcement learning approach for customizing human-robot interfaces, *Proceedingsc of the 13th international conference on Intelligent user interfaces - IUI '09* (2008).

[185] 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.

[186] C.-Y. Chi, R.T.-H. Tsai, J.-Y. Lai and J.Y.-j. Hsu, A Reinforcement Learning Approach to Emotion-based Automatic Playlist Generation, *2010 International Conference on Technologies and Applications of Artificial Intelligence* (2010).

[187] A. Durand and J. Pineau, Adaptive treatment allocation using sub-sampled gaussian processes, in: *2015 AAAI Fall Symposium Series*, 2015.

[188] B. Fernandez-Gauna and M. Grana, Recipe tuning by reinforcement learning in the SandS ecosystem, *2014 6th International Conference on Computational Aspects of Social Networks* (2014).

[189] 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.

[190] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, Exploiting Reinforcement Learning to Profile Users and Personalize Web Pages, *2014 IEEE 38th International Computer Software and Applications Conference Workshops* (2014).

[191] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, On personalizing Web content through reinforcement learning, *Universal Access in the Information Society* **16**(2) (2016), 395–410.

[192] S. Ferretti, S. Mirri, C. Prandi and P. Salomoni, User centered and context dependent personalization through experiential transcoding, *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)* (2014).

[193] L. Fournier, Learning capabilities for improving automatic transmission control, *Proceedings of the Intelligent Vehicles '94 Symposium* (1994).

[194] D. Glowacka, T. Ruotsalo, K. Konuyshkova, k. Athukorala, S. Kaski and G. Jacucci, Directing exploratory search, *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13* (2013).

[195] Y. Goldberg and M.R. Kosorok, Q-learning with censored data, *The Annals of Statistics* **40**(1) (2012), 529–560.

[196] J. Hemminghaus and S. Kopp, Adaptive Behavior Generation for Child-Robot Interaction, *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18* (2018).

[197] T. Hiraoka, G. Neubig, S. Sakti, T. Toda and S. Nakamura, Learning cooperative persuasive dialogue policies using framing, *Speech Communication* **84** (2016), 83–96.

[198] A.S. Lan and R.G. Baraniuk, A Contextual Bandits Framework for Personalized Learning Action Selection, in: *EDM*, 2016.

[199] E. Liebman and P. Stone, DJ-MC: A Reinforcement-Learning Agent for Music Playlist Recommendation, in: *AAMAS*, 2015.

[200] Llorente and S.E. Guerrero, Increasing Retrieval Quality in Conversational Recommenders, *IEEE Transactions on Knowledge and Data Engineering* **24**(10) (2012), 1876–1888.

[201] T. Mahmood, G. Mujtaba and A. Venturini, Dynamic personalization in conversational recommender systems, *Information Systems and e-Business Management* **12**(2) (2013), 213–238.

[202] D. Neumann, T. Mansi, L. Itu, B. Georgescu, E. Kayvanpour, F. Sedaghat-Hamedani, J. Haas, H. Katus, B. Meder, S. Steidl and et al., Vito – A Generic Agent for Multi-physics Model Personalization: Application to Heart Modeling, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (2015), 442–449.

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

[204] C. Peng and P. Vuorimaa, Automatic Navigation Among Mobile DTV Services, in: *ICEIS*, 2004.

[205] C. Peng and P. Vuorimaa, Automatic Navigation Among Mobile DTV Services, in: *ICEIS*, 2004.

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

[207] 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.

[208] D. Shawky and A. Badawi, A Reinforcement Learning-Based Adaptive Learning System, *Advances in Intelligent Systems and Computing* (2018), 221–231.

[209] 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.

[210] A. Srivihok and P. Sukonmanee, Intelligent Agent for e-Tourism: Personalization Travel Support Agent using Reinforcement Learning, in: *WWW 2005*, 2005.

[211] K. Tsiakas, M. Huber and F. Makedon, A multimodal adaptive session manager for physical rehabilitation exercising, *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '15* (2015).

[212] K. Tsiakas, M. Papakostas, B. Chebaa, D. Ebert, V. Karkaletsis and F. Makedon, An Interactive Learning and Adaptation Framework for Adaptive Robot Assisted Therapy, *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '16* (2016).

[213] 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, *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '17* (2017).

[214] G. Vasan and P.M. Pilarski, Learning from demonstration: Teaching a myoelectric prosthesis with an intact limb via reinforcement learning, *2017 International Conference on Rehabilitation Robotics (ICORR)* (2017).

[215] L. Wang, Y. Gao, C. Cao and L. Wang, Towards a General Supporting Framework for Self-Adaptive Software Systems, *2012 IEEE 36th Annual Computer Software and Applications Conference Workshops* (2012).

[216] X. Wang, Y. Wang, D. Hsu and Y. Wang, Exploration in Interactive Personalized Music Recommendation, *ACM Transactions on Multimedia Computing, Communications, and Applications* **11**(1) (2014), 1–22.

[217] S.-T. Yuan, A personalized and integrative comparison-shopping engine and its applications, *Decision Support Systems* **34**(2) (2003), 139–156.

[218] S. Zaidenberg and P. Reignier, Reinforcement Learning of User Preferences for a Ubiquitous Personal Assistant, in: *Advances in Reinforcement Learning*, IntechOpen, 2011.

[219] S. Zaidenberg, P. Reignier and J.L. Crowley, Reinforcement Learning of Context Models for a Ubiquitous Personal Assistant, *3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008* (2008), 254–264.

[220] T. Zhao and I. King, Locality-Sensitive Linear Bandit Model for Online Social Recommendation, *Lecture Notes in Computer Science* (2016), 80–90.

[221] S. Ávila-Sansores, F. Orihuela-Espina and L. Enrique-Sucar, Patient Tailored Virtual Rehabilitation, *Biosystems & Biorobotics* (2013), 879–883.

[222] O. Al-Ubaydli and P.A. McLaughlin, RegData: A numerical database on industry-specific regulations for all United States industries and federal regulations, 1997–2012, *Regulation & Governance* **11**(1) (2017), 109–123, RegGov-10-2014-0122.R3.

[223] A. Artosi, G. Governatori and G. Sartor, Towards a Computational Treatment of Deontic Defeasibility., in: *DEON*, Springer, 1996, pp. 27–46.

[224] T.J. Bench-Capon and F.P. Coenen, Isomorphism and legal knowledge based systems, *Artificial Intelligence and Law* **1**(1) (1992), 65–86.

[225] D. Borsa, B. Piot, R. Munos and O. Pietquin, Observational Learning by Reinforcement Learning, *arXiv preprint arXiv:1706.06617* (2017).

[226] M.K. Bothe, L. Dickens, K. Reichel, A. Tellmann, B. Ellger, M. Westphal and A.A. Faisal, The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas, *Expert review of medical devices* **10**(5) (2013), 661–673.

[227] C. Boutilier, A POMDP formulation of preference elicitation problems, in: *AAAI/IAAI*, 2002, pp. 239–246.

[228] T.D. Breaux, *Legal requirements acquisition for the specification of legally compliant information systems*, North Carolina State University, 2009.

[229] M. Broy, B. Jonsson, J.-P. Katoen, M. Leucker and A. Pretschner, *Model-based testing of reactive systems: advanced lectures*, Vol. 3472, Springer, 2005.

[230] P. Brusilovsky, Methods and techniques of adaptive hypermedia, *User modeling and user-adapted interaction* **6**(2–3) (1996), 87–129.

[231] P.F. Castro and T. Maibaum, Deontic logic, contrary to duty reasoning and fault tolerance, *Electronic Notes in Theoretical Computer Science* **258**(2) (2009), 17–34.

[232] B. Chakraborty and S.A. Murphy, Dynamic treatment regimes, *Annual review of statistics and its application* **1** (2014), 447–464.

[233] M.E. Chamie and B. Acikmese, Finite-Horizon Markov Decision Processes with State Constraints, *arXiv preprint arXiv:1507.01585* (2015).

[234] P. Cuijpers, C.F. Reynolds, T. Donker, J. Li, G. Andersson and A. Beekman, Personalized treatment of adult depression: medication, psychotherapy, or both? A systematic review, *Depression and anxiety* **29**(10) (2012), 855–864.

[235] M. Fatemi, L.E. Asri, H. Schulz, J. He and K. Suleman, Policy networks with two-stage training for dialogue systems, *arXiv preprint arXiv:1606.03152* (2016).

[236] S. Fenech, G.J. Pace and G. Schneider, Automatic conflict detection on contracts, in: *International Colloquium on Theoretical Aspects of Computing*, Springer, 2009, pp. 200–214.

[237] G. Fischer, User modeling in human–computer interaction, *User modeling and user-adapted interaction* **11**(1) (2001), 65–86.

[238] M. Fisher, An introduction to executable temporal logics, *The Knowledge Engineering Review* **11**(1) (1996), 43–56.

[239] M. Gašić, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis and S. Young, On-line policy optimisation of bayesian spoken dialogue systems via human interaction, in: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, 2013, pp. 8367–8371.

[240] T. Gordon, G. Governatori and A. Rotolo, Rules and norms: Requirements for rule interchange languages in the legal domain, *Rule interchange and applications* (2009), 282–296.
[241] 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, p. 27.
[242] M.A. Hamburg and F.S. Collins, The path to personalized medicine, *N Engl J Med* **2010**(363) (2010), 301–304.
[243] A. Hans, D. Schneegaß, A.M. Schäfer and S. Udluft, Safe exploration for reinforcement learning., in: *ESANN*, 2008, pp. 143–148.
[244] M. Hauskrecht and H. Fraser, Planning treatment of ischemic heart disease with partially observable Markov decision processes, *Artificial Intelligence in Medicine* **18**(3) (2000), 221–244.
[245] H. Hirsh, C. Basu and B.D. Davison, Learning to personalize, *Communications of the ACM* **43**(8) (2000), 102–106.
[246] 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.
[247] I. Horrocks, U. Hustadt, U. Sattler and R. Schmidt, 4 Computational modal logic, *Studies in Logic and Practical Reasoning* **3** (2007), 181–245.
[248] E. Horvitz, J. Breese, D. Heckerman, D. Hovel and K. Rommelse, The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users, in: *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1998, pp. 256–265.
[249] C. Hu, W.S. Lovejoy and S.L. Shafer, Comparison of some suboptimal control policies in medical drug therapy, *Operations Research* **44**(5) (1996), 696–709.
[250] K. Janowicz, F. Van Harmelen, J.A. Hendler and P. Hitzler, Why the data train needs semantic rails, *AI Magazine* (2014).
[251] 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.
[252] Y. Kadota, M. Kurano and M. Yasuda, Discounted Markov decision processes with utility constraints, *Computers & Mathematics with Applications* **51**(2) (2006), 279–284.
[253] S. Keizer, S. Rossignol, S. Chandramohan and O. Pietquin, Conversational interfaces, in: *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, Springer, 2012, pp. 39–74.
[254] G. Konidaris, L.P. Kaelbling and T. Lozano-Perez, Symbol acquisition for probabilistic high-level planning, *Image (a1, Z0)* **1** (2015), Z0.
[255] 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.
[256] S. Lautenschläger, Walled off? Banking regulation after the crisis, 2017, Speech by Sabine Lautenschläger, Member of the Executive Board of the ECB and Vice-Chair of the Supervisory Board of the ECB, at the Institute of International and European Affairs, Dublin, 13 March 2017 [Accessed: 2017 09 15].
[257] O. Lemon, Conversational interfaces, in: *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, Springer, 2012, pp. 1–4.
[258] Y. Liu, S. Muller and K. Xu, A static compliance-checking framework for business process models, *IBM Systems Journal* **46**(2) (2007), 335–361.
[259] U. Manber, A. Patel and J. Robison, Experience with personalization of Yahoo!, *Communications of the ACM* **43**(8) (2000), 35–39.
[260] J. Mariani, S. Rosset, M. Garnier-Rizet and L. Devillers, Natural Interaction with Robots, Knowbots and Smartphones, *New York* (2014).
[261] J. McCarthy, Phenomenal data mining, *Communications of the ACM* **43**(8) (2000), 75–79.
[262] M. Minsky, Commonsense-based interfaces, *Communications of the ACM* **43**(8) (2000), 66–73.
[263] T.M. Moldovan and P. Abbeel, Safe exploration in Markov decision processes, *arXiv preprint arXiv:1205.4810* (2012).
[264] S.A. Murphy, An experimental design for the development of adaptive treatment strategies, *Statistics in medicine* **24**(10) (2005), 1455–1481.
[265] S.A. Murphy, K.G. Lynch, D. Oslin, J.R. McKay and T. TenHave, Developing adaptive treatment strategies in substance abuse research, *Drug and alcohol dependence* **88** (2007), S24–S30.
[266] L.A. Nguyen, The modal logic programming system MProlog, in: *European Workshop on Logics in Artificial Intelligence*, Springer, 2004, pp. 266–278.
[267] D.W. Oard, J. Kim et al., Implicit feedback for recommender systems, in: *Proceedings of the AAAI workshop on recommender systems*, Menlo Park, CA: AAAI Press, 1998, pp. 81–83.
[268] M. Oishi, C.J. Tomlin, V. Gopal and D. Godbole, Addressing multiobjective control: Safety and performance through constrained optimization, *Lecture notes in computer science* **2034** (2001), 459–472.
[269] E.P. Pednault, Representation is everything, *Communications of the ACM* **43**(8) (2000), 80–80.
[270] T.J. Perkins and A.G. Barto, Lyapunov design for safe reinforcement learning, *Journal of Machine Learning Research* **3**(Dec) (2002), 803–832.

[271] 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.
[272] N. Piterman and A. Pnueli, Synthesis of reactive (1) designs, Springer.
[273] T. Raafat, E. Plettenberg and N. Trokanas,
NextAngles: A Semantic Platform Reimagining Compliance.
[274] O. Ritchie and K. Thompson, The UNIX time-sharing system, *The Bell System Technical Journal* **57**(6) (1978), 1905–1929.
[275] N. Roy, G. Gordon and S. Thrun, Finding approximate POMDP solutions through belief compression, *Journal of artificial intelligence research* **23** (2005), 1–40.
[276] 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.
[277] S. Russell, P. Norvig and A. Intelligence, A modern approach, *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* **25** (1995), 4.
[278] A. Santoro, D. Raposo, D.G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia and T. Lillicrap, A simple neural network module for relational reasoning, *arXiv preprint arXiv:1706.01427* (2017).
[279] A.J. Schaefer, M.D. Bailey, S.M. Shechter and M.S. Roberts, Modeling medical treatment using Markov decision processes, in: *Operations research and health care*, Springer, 2005, pp. 593–612.
[280] B. Schlich, Winning through customer experience, *EY Global Consumer Banking Survey* (2014).
[281] B.a. Schlich, The customer takes control, *EY Global Consumer Banking Survey* (2012).
[282] B. Schlich, D. Epstein, T. SchrezenMaier and A.S. Turner, The relevance challenge: What retail banks must do to remain in the game, *EY Global Consumer Banking Survey* (2016).
[283] S. Shalev-Shwartz, S. Shammah and A. Shashua, Safe, multi-agent, reinforcement learning for autonomous driving, *arXiv preprint arXiv:1610.03295* (2016).
[284] 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.
[285] 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.
[286] D.C. Smith, Building personal tools by programming, *Communications of the ACM* **43**(8) (2000), 92–95.
[287] A. Srivihok and P. Sukonmanee, E-commerce intelligent agent: personalization travel support agent using Q Learning, in: *Proceedings of the 7th international conference on Electronic commerce*, ACM, 2005, pp. 287–292.
[288] G. Tesauro, Temporal difference learning and TD-Gammon, *Communications of the ACM* **38**(3) (1995), 58–68.
[289] P. Thomas and E. Brunskill, Data-efficient off-policy policy evaluation for reinforcement learning, in: *International Conference on Machine Learning*, 2016, pp. 2139–2148.
[290] P.S. Thomas, Safe reinforcement learning (2015).
[291] P.S. Thomas, B.C. da Silva, A.G. Barto and E. Brunskill, On Ensuring that Intelligent Machines Are Well-Behaved, *arXiv preprint arXiv:1708.05448* (2017).
[292] W. Van der Aalst, A. Adriansyah and B. van Dongen, Replaying history on process models for conformance checking and performance analysis, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(2) (2012), 182–192.
[293] R. van Doesburg,
A Formal Method for Interpretation of Sources of Norms.
[294] P. Van Wesel and A.E. Goodloe, Challenges in the Verification of Reinforcement Learning Algorithms (2017).
[295] K.R. Varshney and H. Alemzadeh, On the safety of machine learning: Cyber-physical systems, decision sciences, and data products, *Big Data* **5**(3) (2017), 246–255.
[296] A.Z. Wyner, A functional program for agents, actions, and deontic specifications, in: *International Workshop on Declarative Agent Languages and Technologies*, Springer, 2006, pp. 239–256.
[297] V. Zamborlini, R. Hoekstra, M.d. Silveira, C. Pruski, A. ten Teije et al., Inferring recommendation interactions in clinical guidelines: case-studies on multimorbidity (2015).
[298] 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.