

Knowledge-Driven Adaptive Interaction in a Healthcare Dialogue Agent

Efthymios GRIGORAKIS^a, Shihan WANG^a, Maaïke DE BOER^{b,1}

^a*Information and Computing Sciences, Utrecht University, the Netherlands*

^b*dep Data Science, TNO, the Netherlands*

Abstract. The rising percentages of diabetes patients continue to stress the global healthcare system. As general practitioners do not have the time to monitor all patients, hybrid human-AI interaction could provide a solution. We propose a knowledge-based dialogue agent that leverages Reinforcement Learning (RL) and Knowledge Graphs (KG) to achieve adaptive interactions. In this paper, we use a hierarchical architecture, decomposing the problem into a master component and two specialized workers: a knowledge acquisition worker gathering information about the user and a recommendation worker providing lifestyle change recommendations to patients. The two workers are driven by two KGs: a medical graph and a user graph. Our proposed method uses a Relational Graph Attention Network (RGAT) Encoder and a Cross-Graph Attention module to encode and fuse the information included in both graphs. Based on experimentation on both a state-of-the-art dataset (MovieLens) and a synthetic diabetes dataset, we show that the master component learns an effective policy for selecting the right worker. Future work includes benchmarking our approach against real-world patient data, and integrating the method in a real application to verify real world performance.

Keywords. Reinforcement Learning, Knowledge Graphs, Human-AI Interaction, Diabetes Care, Healthcare Dialogue Agent

1. Introduction

Nowadays, the Diabetes disease, which is a disorder of glucose metabolism, is considered one of the most common diseases around the world and rapidly increasing in number of cases [1,2]. Patients must be equipped with self-management skills in order to overcome daily difficulties with Diabetes and maintain a normal and healthy daily routine. The complex nature of Diabetes, combined with the continuous communication between the patient and a group of experts regarding life changes, resulted in understanding the technology-centered viewpoint of current Diabetes management products. That problem could be tackled by using Hybrid Intelligence [3,4], although the adoption of the HI philosophy in supporting lifestyle changes has introduced new challenges, as it requires the collaboration among experts from different research fields, lacks clear evaluation baselines and raises numerous ethical concerns [5].

In this paper, we build upon a proposed interactive system to help people with Type 2 Diabetes adopt a healthier lifestyle [6,7]. This system has five modules: 1) a dialogue

¹Corresponding Author: Maaïke de Boer, maaïke.deboer@tno.nl

component, that is based on a rule-based approach to gather information about the patient; 2) an Information Extraction module, which transforms text into structured triples, that can capture the context of the dialogue [8]; 3) a dynamic User Knowledge Graph (User KG), which represents the user's preferences based on the triples and will be used for personalized recommendations; 4) a static Domain Knowledge Graph (Domain KG) contains medical information about the Diabetes disease; 5) a Reasoning Engine that leverages both the User KG and the Domain KG, in order to select the most appropriate action. This work focuses on leveraging Reinforcement Learning and Knowledge Graphs for the last step. In particular, we propose to combine a master component and two specialized *worker* agents: a knowledge acquisition worker gathering information about the user and a recommendation worker providing lifestyle change recommendations to patients. The two workers are driven by two KGs: a medical graph and a user graph respectively. We evaluated our approach in an open-source dataset (MovieLens) and a synthetic diabetes dataset. Through this preliminary study, we aim to demonstrate a proof of concept for integrating structured knowledge and learning-based dialogue policies in the healthcare domain, paving the way for future real-world data studies.

2. Related Work

Recently, the integration of Reinforcement Learning (RL) with Knowledge Graphs in dialogue systems has attracted significant attention in various research fields [9,10]. In our focused healthcare field, RL tends to be used for providing dynamic treatment to patients [11], automated disease diagnosis [12,13,14] and general medical fields such as resource allocation and drug discovery [11]. We have been inspired by several works on knowledge-enhanced Hierarchical RL [12,15,14], in which low-level Deep Q-Network agents are for example associated with a subset of diseases and a high-level agent chooses over the different low-level agents. For instance, in [15] a relation matrix and a medical knowledge graph are added to capture relations between diseases and symptoms.

In the field of recommendations, RL agents are used to either suggest an item or ask for the user preferences as attributes [16], in multitask recommendation using an actor-critic framework [17] and using hierarchical RL [18]. In Question Answering dialogues, [19] combined Hierarchical Reinforcement Learning and Large Language Models (LLMs), to create an effective question answering system. [20] creates an agent that can acquire and augment its own knowledge, by interacting through a conversation.

Building on these lines of work, we argue for the potential of adopting RL-based approaches in a hierarchical manner [21], for jointly achieving question-asking (to enrich knowledge) and recommendation (based on knowledge) within one unified dialogue system.

3. Methodology

We propose a dialogue agent that leverages Reinforcement Learning and Knowledge Graphs. By leveraging RL, we can capture personal preferences through feedback from the environment, which will help us make personalized and accurate suggestions. On the other hand, the Knowledge Graphs give us the ability to organize knowledge in a

structured way, capture semantics between entities, and let the agent reason over user's and domain's structured knowledge. The dialogue agent can have a conversation with the patient, in order to acquire more knowledge about the patient's personal and medical history, as well as knowledge about clinical measurements that reflect the patient's state. Combining this information with domain knowledge that can be found in the Medical Knowledge Graph, the agent can suggest changes to the patient.

We are inspired by the hierarchical RL setup, in particular the sub-class of policy tree methods introduced in [21]. We propose to decompose the problem into subtasks, learning a high-level policy to select the corresponding subtask and allowing each subtask to have its own separate policy in the lower-level. Our system consists of 3 components: 1) Master: Select the appropriate low-level worker; 2) Knowledge Acquisition (KA) Worker: Ask questions and acquire information about the user; 3) Recommendation Worker: Provides recommendations based on the information about the user. Notably, all components in our system share a common objective: to provide appropriate personalized recommendations based on structured knowledge.

3.1. Master

The goal of the master component is to learn to select between the low-level workers, that interact with the user. We will represent the main task of personalized recommendations as Γ and the two subtasks as ω_1 for the knowledge acquisition task and ω_2 for the recommendation task. The main task policy is notated as π_{θ_T} . We adopt an Actor-Critic architecture for the master. **Actor** (π_{θ_T}) determines the probability of selecting knowledge acquisition or recommendation worker. And **Critic** (V_{θ_T}) estimates the value of the current state to critique the Actor's choice.

We augment the master state $\mathbf{s}_{master}$ with the entropy of the recommendation worker $\mathcal{H}_{rec}$ and t_{index} the turn of the conversation. $\mathcal{H}_{rec}$ will allow the master agent understand when the recommendation worker is certain about the recommendation that will provide or needs more information about the user. Since we have a turn-based conversation, turn t_{index} provides the information about how long the conversation will still exist with the user. Thus, the final master state at a timestep t is $\mathbf{x}_t = [\mathbf{s}_{master} || \mathcal{H} || t_{index}]$, where $||$ the operation of concatenation and the state representation $\mathbf{s}_{master}$ is generated by user encoder in Sect.3.4. The Actor part of the master agent uses the master state via Eq. 1-3. The Critic network represents a state value function, estimating the expected future returns from that state (4-6).

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1 \mathbf{x}_t) + \mathbf{b}_1 \quad (1) \quad \mathbf{h}'_1 = \text{ReLU}(\mathbf{W}'_1 \mathbf{x}_t) + \mathbf{b}'_1 \quad (4)$$

$$\mathbf{h}_2 = \text{ReLU}(\mathbf{W}_2 \mathbf{h}_1) + \mathbf{b}_2 \quad (2) \quad \mathbf{h}'_2 = \text{ReLU}(\mathbf{W}'_2 \mathbf{h}_1) + \mathbf{b}'_2 \quad (5)$$

$$\mathbf{z} = \mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3 \quad (3) \quad V(\mathbf{x}_t) = \mathbf{W}_3' \mathbf{h}'_2 + \mathbf{b}_3 \quad (6)$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ are the weights of the first hidden, second hidden and output layer, respectively. We use a Softmax layer to calculate the policy $\pi_{\theta_T}(a|\mathbf{x}_t)$.

Since our problem is a turn-based problem with a maximum number of turns that the agent can interact with the user, we would like the reward function that will allow the master agent to maximize its reward by finding the minimum sequence of interactions required to provide the appropriate recommendation for the user. For that reason, we define the **reward** r_t^Γ for the master agent as -1 if KA worker is selected, -1 if Rec

Worker is selected and ground truth not found. When Ground truth is found, we set +10 positive reward and distribute this reward uniformly in the actions among the selected KA workers and the final recommendation worker. In ablation studies, different reward values are tested of which this setting is the best.

3.2. Knowledge Acquisition (KA) Worker

The first low-level worker, starting from an empty user knowledge graph (KG) G^{user} , focuses on acquiring information relevant to the user's preferences and incorporates it into this KG. The actions that can be selected from that worker are entities from the domain knowledge graph that are included in the set of attributes. The final action for the worker consists of two components: concept (c) is the type of the question and attribute (v) is the specific entity from the selected concept. The policy of the KA worker can be written as:

$$\pi_{KA}(c, v | \mathbf{s}_{ka}) = P(c | \mathbf{s}_{ka}) \cdot P(v | \mathbf{s}_{ka}, c) \quad (7)$$

where $\mathbf{s}_{ka}$ is a representation of the KA worker state.

We adopt a similar technical design as the Master component. The state $\mathbf{s}_{ka}$ is processed by a shared backbone that follows the same architecture as in Eq. 1, 2, 3. Given the backbone output, the agent first decides which type of information will acquire in that round by selecting the concept c and attribute v based on the softmax function. The Critic network will estimate the value $V(\mathbf{s}_{ka})$ of the state, utilizing the same architecture as the Critic network in the master. Given the backbone output $\mathbf{h}_{ka}$, the agent first decides which type of information will acquire in that round by selecting the concept c and attribute v based on:

$$\begin{aligned} P(c | \mathbf{s}_{ka}) &= \text{Softmax}(\mathbf{W}_C \mathbf{h}_{ka} + \mathbf{b}_C) \\ P(v | \mathbf{s}_{ka}, c) &= \text{Softmax}(\mathbf{W}_{A,C} \mathbf{h}_{ka} + \mathbf{b}_{A,C}) \end{aligned} \quad (8)$$

After the selection of the current action, user will response their preference. The user preference is presented as a triple with $\langle user \rangle \text{ likes/neutral/dislikes } \langle entity \rangle$. Those new information is added in the G^{user} and included in the final embedding of the user KG (see Sect.3.4 for the encoder details). The goal of the KA worker, is to identify the most relevant questions, which will help the recommendation worker to provide personalized recommendations to the user. Thus, we would like our reward to reflect how useful was the acquired information for the recommendation worker:

$$r_t^{\omega_1} = \log(\pi_{rec}(gt | \mathbf{x}_t)) - \log(\pi_{rec}(gt | \mathbf{x}_{t-1})) \quad (9)$$

where gt is the ground truth entity for the current user. In other words, if the obtained information from the KA worker can enhance the knowledge of our user, a positive reward will be given.

3.3. Recommendation (Rec) Worker

This worker is responsible for leveraging the acquired knowledge and providing recommendations to the user. The action space $\mathcal{I}_{rec}$ is a set of recommendation entities, and the state of the recommendation worker is s_{rec} , another output of the user encoder. While we follow the same Actor-Critic structure as the master, the policy of actor network of Rec worker is derived by applying a softmax layer over the scores of all the possible recommended items. When the recommendation worker provides a recommendation to the user, there are two cases: 1) the reward of that worker is +10 if the recommendation is same as the ground truth, and 2) otherwise the rating as given by the user to this recommended entity.

3.4. State Representation

User Encoder To be able to utilize the semantic and structural information the user KG contains, we introduce the RGAT-Encoder to encode KG information into state representations, based on the Relational Graph Attention Network (RGAT) [22]. Each relation r is processed by a specific GAT convolutional layer, allowing the encoder to learn relation-specific weights, passing the message of the relation into the output embeddings of that state. We use the Attentional Global pooling of the final stage. That stage calculates the single graph embeddings by applying a weighted sum. The RGAT-Encoder, except from the gradients of the master and the two workers, is also trained with an auxiliary Contrastive loss. In order to train the RGAT, we use a ComplEx [23] decoder to compute the scores. The embeddings of the knowledge graph are pre-trained and remain frozen during the training procedure of RL.

Cross-Graph Attention Cross-Graph attention is also used to combine the information of user KG (user's vector) and medical domain KG (entities available). This can help the agent understand which recommendations are more relevant to user's preferences. The multihead attention is used for the user vector and a Feed Forward Network with a GELU activation function is used to create the final layer. We adopt a Hard Parameter Sharing architecture [24] and let the User Encoder and the Cross-Graph Attention act as a shared encoder for all three components in the dialogue agent. In this way, we learn a generalized state representation for the master and the workers, while a set of task-specific projection layers optimize this shared representation for each agent's policy.

4. Experimental Setup

Datasets MovieLens is used in the evaluation of other works that utilize RL for a recommendation system [25,26,27,28,29]. The MovieLens dataset exists in multiple versions, all consisting of timestamped user ratings of movies on a 1-5 scale. We selected the MovieLens-2k² dataset, which is an extension of the MovieLens-10M³ dataset, by including information about actors and directors for each movie, rather than only the genre. The dataset consists of 2,113 users, 10,197 movies, 855,598 ratings, 20 genres, 95,321 actors and 4,060 directors.

²<https://grouplens.org/datasets/hetrec-2011/>

³<https://grouplens.org/datasets/movielens/10m/>

A synthetic dataset is created for patient-intervention interactions, based on the MovieLens dataset. Users are mapped to Patients, Movies to Lifestyle Interventions (e.g. running routines, diet plans), Genres to Intervention Category (e.g. Physical Activity, Nutrition), Actors to Clinical Components (specific elements of Interventions) and Directors to Clinical Protocols. To generate the Patient-Intervention interactions, we modeled each patient's profile based on their user historical preference genre patterns. We selected the interactions by uniformly sampling interventions from the corresponding Intervention Category. To generate the connections between the interventions and clinical components and clinical protocols, we learn the cardinality distribution of the corresponding data category in the MovieLens dataset. Then, after randomly sampling the cardinality of a specific attribute for an intervention, we uniformly sample between the values of that attribute. The last step of the synthetic data creation is the generation of the rating for each Patient-Intervention. To generate the latter, we use a linear model with additive noise.

$$r_{p,i} = \mu_{base} + 0.5 * \mu_{\text{Clinical Components}} + 0.5 * \mu_{\text{Clinical Protocols}} + \varepsilon \quad (10)$$

where $r_{p,i}$ the rating of patient p to intervention i , μ_{base} the base score, which is set to 2.5 that represents the neural opinion since we use a 0 – 5 rating scale, $\mu_{\text{Clinical Components}}$ and $\mu_{\text{Clinical Protocols}}$ the mean score of the attributes that are included in the intervention respectively, and ε Gaussian noise. The dataset consists of 2,113 users, 7,000 interventions, 1,328,639 ratings, 2,971 protocols, 20 categories and 45,876 components.

Domain KG To construct the domain knowledge graph for the MovieLens dataset, each movie will be linked to its actors, directors, and genres via the relationships *hasActor*, *hasDirector*, *hasGenre*, respectively. This results in 776 nodes and 1,261 edges. In a similar procedure, in the domain knowledge graph for the synthetic diabetes dataset, interventions are connected with the protocols, the categories, and the components via the relationships *hasProtocol*, *hasCategory*, *hasComponent*, respectively. This results in 507 nodes and 1,046 edges. A train and test set are created using the items that are included in the Domain Knowledge Graph. We only keep items with a positive interaction rating, since we want our agent to learn to provide successful recommendations to the user. Lastly, the target item for each user is set to the item that satisfies the procedure described above and is the most recent chronologically item.

User Simulator For training and evaluating our RL-based agent in simulations before real-user studies, we implemented a user simulator based on the preference model in [30]. The user simulator captures the personal preference of the user and can response to our dialogue agent at each timestamp. Assuming that the user has interacted with historical items and the user is asked about attribute j . The preference of the user about that attribute is given by:

$$r_j = \frac{1}{|I_j|} \sum_{i \in I_j} r_i \quad (11)$$

where I_j denotes historical items with that attribute j and r_i the rating of the item i , with r_i being normalized and $r_i \in [-1, 1]$.

Baselines Following the recommendation systems community, we use a Matrix Factorization model [31] as our baseline model. This model projects the users and the items in a shared d -dimension latent space. The predicted rating is computed by the dot product between the user embedding (from the user KG) and the item embedding (from the domain KG). This baseline model uses only pure latent interactions without any other features or nonlinearities. Additionally, to capture non-linear interactions, we also use the Neural Matrix Factorization [32]. In this model, a Generalized Matrix Factorization (GMF) component, which generalizes the dot product, is fused with a Multi-Layer Perceptron (MLP) that learns a non-linear interaction function.

Parameters All experiments were executed on an NVIDIA L40S GPU. The learning rate was selected as $\eta = 10^{-4}$, while the embedding dimension $d_{emb} = 256$. The master was trained with the PPO algorithm [33], with 5 epochs and a clipping parameter $\epsilon = 0.2$. The low-level workers were trained using the REINFORCE algorithm [34]. All the presented results are averaged across three simulation runs.

Evaluation Metrics To evaluate the recommendation performance of the models, we selected the Hits@K metric, denoting how often the ground truth recommendation is within the top-k recommendations of the model, as used in the recommendation community [32,17]. In our study, we selected the Hits@20 metric.

5. Results & Discussion

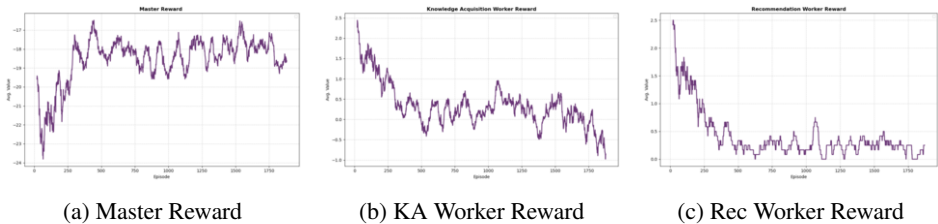


Figure 1. Synthetic Diabetes Dataset training



Figure 2. MovieLens Dataset training

Figure 1 and 2 show the rewards for the two datasets during training of the RL-based dialogue agent. Two example conversations are presented in Fig. 3. Table 1 presents the results of the comparison of the recommendation performance between our trained agent and the baselines. The results show that our learned RL and knowledge-leveraged agent

Model	Datasets	
	MovieLens	Synthetic Diabetes
Our Approach	0.39 ($\pm$ 0.01)	0.33 ($\pm$ 0.02)
Matrix Factorization (baseline)	0.22 ($\pm$ 0.02)	0.28 ($\pm$ 0.03)
Neural Matrix Factorization (baseline)	0.11 ($\pm$ 0.01)	0.14 ($\pm$ 0.01)

Table 1. Hits@20 scores for our approach compared to the baseline matrix factorization methods

System: You should try Intervention.1891. User: I am neutral about Intervention.1891. System: Do you like Protocol.517? User: I am neutral about Protocol.517. System: You should try Intervention.5814. User: I am neutral about Intervention.5814. System: Do you like Category.12? User: Yes, I like Category.12. System: You should try Intervention.6931. User: Yes, I like Intervention.6931. System: You should try Intervention.4373. User: I am neutral about Intervention.4373. System: You should try Intervention.3740. User: I am neutral about Intervention.3740. System: You should try Intervention.4935. User: No, I dislike Intervention.4935. System: Do you like Category.8? User: No, I dislike Category.8. System: You should try Intervention.762. User: I am neutral about Intervention.762.	System: Do you like Category.4? User: No, I dislike Category.4. System: Do you like Category.17? User: Yes, I like Category.17. System: Do you like Protocol.836? User: Yes, I like Protocol.836. System: Do you like Protocol.2989? User: I am neutral about Protocol.2989. System: You should try Intervention.3145. User: I am neutral about Intervention.3145. System: Do you like Protocol.79? User: No, I dislike Protocol.79. System: You should try Intervention.6239. User: I am neutral about Intervention.6239. System: Do you like Protocol.944? User: Yes, I like Protocol.944. System: Do you like Component.74? User: I am neutral about Component.74. System: You should try Intervention.6931. User: Yes, I like Intervention.6931.
---	---

Figure 3. Synthetic Diabetes dialogues at early (left) and late (right) training phases

performs better than both baselines in providing recommendations. During the learning process, the master reward behaves as expected across both datasets, whereas the low-level workers exhibit different patterns. In the MovieLens dataset, the KA and Rec workers show delayed learning, with positive performance emerging only at later stages. In contrast, in the synthetic diabetes dataset, the rewards of both low-level workers decrease over time. Further ablation studies reveal that training either the KA worker or the Rec worker individually leads to stable and strong performance. This suggests that jointly training both workers introduces non-stationarity, which may result in an unstable learning process. To further investigate this issue, we plan to evaluate the approach on a third dataset. We also perform experiments to evaluate Gross-Graph attention design and the results show our design for combining two KGs improves the recommendation performance.

6. Conclusion and Future Work

In this work, we proposed a knowledge-based dialogue agent that leverages Reinforcement Learning and Knowledge Graphs to achieve adaptive interactions. From the experiments conducted, we saw that the master, responsible for selecting between the low-level workers and guiding the conversation, was able to learn an effective policy for its task. While our approach shows promising results and outperforms baseline recommendation models, it has so far only been evaluated on synthetic data, and further investigation of the low-level workers is required. In future work, we aim to perform further analysis, benchmarking our approach against complex real-world patient conversations to gain useful insights into our system’s performance. In particular, we plan to evaluate the capability and effectiveness of our dialogue agent within real healthcare settings.

References

- [1] Roglic G. WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*. 2016;1(1):3-8.
- [2] Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*. 2021 12;183:109119. Available from: <https://doi.org/10.1016/j.diabres.2021.109119>.
- [3] de Boer MH, van der Waa J, van Gent S, Smit QT, Korteling W, van Stokkum RM, et al. A contextual hybrid intelligent system design for diabetes lifestyle management. In: *International Workshop Modelling and Representing Context, ECAI*. vol. 23; 2023. .
- [4] Dudzik BJW, Van Der Waa JS, Chen PY, Dobbe R, De Troya IMDR, Bakker RM, et al. ViewPoint: Hybrid Intelligence supports application development for diabetes lifestyle management. *Journal of Artificial Intelligence Research*. 2024 7;80:919-29. Available from: <https://doi.org/10.1613/jair.1.15916>.
- [5] Hao C, Uusitalo S, Figueroa C, Smit QTS, Strange M, Chang WT, et al. A human-centered perspective on research challenges for hybrid human artificial intelligence in lifestyle and behavior change support. *Frontiers in Digital Health*. 2025 3;7. Available from: <https://doi.org/10.3389/fdgth.2025.1544185>.
- [6] Chen PY, Baez Santamaria S, de Boer MH, den Hengst F, Kamphorst BA, Smit Q, et al. Intelligent Support Systems for Lifestyle Change: Integrating Dialogue, Information Extraction, and Reasoning. In: *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press; 2024. p. 457-9.
- [7] den Hengst F, Alaka S, Kamphorst BA. Collaborative hybrid intelligence platform CHIP: A modular architecture for developing and testing personalized lifestyle support interactions. *SoftwareX*. 2026;33:102536.
- [8] Ntanavaras S. *Conversational Triple Extraction for Diabetes Healthcare Management Using Synthetic Data*. Vrije Universiteit Amsterdam; 2024.
- [9] Huo Q, Fu H, Song C, Sun Q, Xu P, Qu K, et al. Knowledge Graph Based on Reinforcement Learning: A survey and New Perspectives. *IEEE Access*. 2024 1;12:161897-924. Available from: <https://doi.org/10.1109/access.2024.3479774>.
- [10] Liu Y, Wang H, Zhou H, Li M, Hou Y, Zhou S, et al. A review of reinforcement learning for natural language processing and applications in healthcare. *Journal of the American Medical Informatics Association*. 2024 8;31(10):2379-93. Available from: <https://doi.org/10.1093/jamia/ocae215>.
- [11] Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*. 2021;55(1):1-36.
- [12] Zhu Y, Li Y, Cui Y, Zhang T, Wang D, Zhang Y, et al. A Knowledge-Enhanced Hierarchical Reinforcement Learning-Based dialogue system for automatic disease diagnosis. *Electronics*. 2023 12;12(24):4896. Available from: <https://doi.org/10.3390/electronics12244896>.
- [13] Huang Y, Chen M, Tang K. Training like Playing: A Reinforcement Learning And Knowledge Graph-based framework for building Automatic Consultation System in Medical Field. *arXiv e-prints*. 2021 Jun:arXiv:2106.07502.
- [14] Tiwari A, Saha S, Bhattacharyya P. A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning. *Knowledge-Based Systems*. 2022 1;242:108292. Available from: <https://doi.org/10.1016/j.knosys.2022.108292>.
- [15] Xu L, Zhou Q, Gong K, Liang X, Tang J, Lin L. End-to-End Knowledge-Routed Relational Dialogue System for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019 7;33(01):7346-53. Available from: <https://doi.org/10.1609/aaai.v33i01.33017346>.
- [16] Deng Y, Li Y, Sun F, Ding B, Lam W. Unified Conversational recommendation policy learning via graph-based reinforcement learning. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021 7. Available from: <https://doi.org/10.1145/3404835.3462913>.
- [17] Chen K, Sun S. Knowledge-Based Conversational Recommender Systems Enhanced by Dialogue Policy Learning. In: *Proceedings of the 10th International Joint Conference on Knowledge Graphs, IJCKG '21*. New York, NY, USA: Association for Computing Machinery; 2022. p. 10–18. Available from: <https://doi-org.utrechtuniversity.idm.oclc.org/10.1145/3502223.3502225>.
- [18] Zhang M, Li Y, Li S, Wang Y, Yan J. Knowledge Graph-enhanced hierarchical reinforcement learning for interactive and explainable recommendation. *IEEE Access*. 2024 1;12:137345-59. Available from:

- <https://doi.org/10.1109/access.2024.3464733>.
- [19] Zhang Z, Zhao W. A Collaborative Reasoning Framework Powered by Reinforcement Learning and Large Language Models for Complex Questions Answering over Knowledge Graph. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics; 2025. p. 10672-84. Available from: <https://aclanthology.org/2025.coling-main.712/>.
 - [20] Santamaria SB, Wang S, Vossen P. Knowledge acquisition for dialogue agents using reinforcement learning on graph representations; 2024. Available from: <https://arxiv.org/abs/2406.19500>.
 - [21] Pateria S, Subagdja B, Tan Ah, Quek C. Hierarchical Reinforcement Learning: A Comprehensive Survey. ACM Comput Surv. 2021 Jun;54(5). Available from: <https://doi.org.utrechtuniversity.idm.oclc.org/10.1145/3453160>.
 - [22] Busbridge D, Sherburn D, Cavallo P, Hammerla NY. Relational graph attention networks. arXiv preprint arXiv:190405811. 2019.
 - [23] Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex Embeddings for Simple Link Prediction. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning. vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA: PMLR; 2016. p. 2071-80. Available from: <https://proceedings.mlr.press/v48/trouillon16.html>.
 - [24] Ruder S. An Overview of Multi-Task Learning in Deep Neural Networks. CoRR. 2017;abs/1706.05098. Available from: <http://arxiv.org/abs/1706.05098>.
 - [25] Zhao X, Zhang W, Wang J. Interactive collaborative filtering. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. CIKM '13. New York, NY, USA: Association for Computing Machinery; 2013. p. 1411–1420. Available from: <https://doi.org/10.1145/2505515.2505690>.
 - [26] Zhao Z, Chen X, Xu Z, Cao L. Tag-Aware Recommender System Based on Deep Reinforcement Learning. Mathematical Problems in Engineering. 2021;2021(1):5564234. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/5564234>.
 - [27] Izadkhan S, Reakhdar B. Deep Reinforcement Learning Based Group Recommendation System with Multi-Head Attention Mechanism. In: 2023 Fifth International Conference on Transdisciplinary AI (TransAI); 2023. p. 120-7.
 - [28] Grishanov A, Ianina A, Vorontsov K. Multiobjective Evaluation of Reinforcement Learning Based Recommender Systems. In: Proceedings of the 16th ACM Conference on Recommender Systems. RecSys '22. New York, NY, USA: Association for Computing Machinery; 2022. p. 622–627. Available from: <https://doi.org/10.1145/3523227.3551485>.
 - [29] Peng S, Siet S, Ilkhomjon S, Kim DY, Park DS. Integration of Deep Reinforcement Learning with Collaborative Filtering for Movie Recommendation Systems. Applied Sciences. 2024;14(3). Available from: <https://www.mdpi.com/2076-3417/14/3/1155>.
 - [30] Zhang S, Balog K. Evaluating Conversational Recommender Systems via User Simulation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 1512–1520. Available from: <https://doi.org/10.1145/3394486.3403202>.
 - [31] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer. 2009;42(8):30-7.
 - [32] He X, Liao L, Zhang H, Nie L, Hu X, Chua TS. Neural Collaborative Filtering; 2017. Available from: <https://arxiv.org/abs/1708.05031>.
 - [33] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal Policy Optimization Algorithms; 2017. Available from: <https://arxiv.org/abs/1707.06347>.
 - [34] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning. 1992;8(3):229-56.