Recommending safe actions by learning from sub-optimal demonstrations

Lars Boecking
boecking@fzi.de
FZI Research Center for Information Technology
Karlsruhe, Germany

Patrick Philipp
philipp@fzi.de
FZI Research Center for Information Technology
Karlsruhe, Germany

ABSTRACT

Clinical pathways describe the treatment procedure for a patient from a medical point of view. Based on the patient's condition, a decision is made about the next actions to be carried out. Such recurring sequential process decisions could well be outsourced to a reinforcement learning agent, but the patient's safety should always be the main consideration when suggesting activities. The development of individual pathways is also cost and time intensive, therefore a smart agent could support and relieve physicians. In addition, not every patient reacts in the same way to a clinical intervention, so the personalization of a clinical pathway should be given attention.
In this paper we address with the fundamental problem that the use of reinforcement learning agents in the specification of clinical pathways should provide an individual optimal proposal within the limits of safety constraints.
Imitating the decisions of physicians can guarantee safety but not optimality. Therefore, we present an approach that ensures compliance with health critical rules without limiting the exploration of the optimum. We evaluate our approach on open source gym environment where we are able to show that our adaptation of behavior cloning not only adheres better to safety regulations, but also manages to better explore the space of the optimum in the collective rewards.

CCS CONCEPTS

• **Applied computing** → **Health care information systems**.

KEYWORDS

health care; clinical pathway; reinforcement learning; personalization; imitation learning ; safety; constraints

ACM Reference Format:
Lars Boecking and Patrick Philipp. 2020. Recommending safe actions by learning from sub-optimal demonstrations. In *Proceedings of the 5th International Workshop on Health Recommender Systems co-located with 14th ACM Conference on Recommender Systems (HealthRecSys'20), Online, Worldwide, September 26, 2020* , 8 pages.

HealthRecSys'20, September 26th, 2020, Online, Worldwide

1 INTRODUCTION

Our work focuses on the use of reinforcement learning to optimize and personalize clinical pathways, illustrated in Figure 1. Rehabilitation procedure, called „**Clinical Pathway**", describes in detail which activities are to be carried out for a patient within a course of treatment[13].

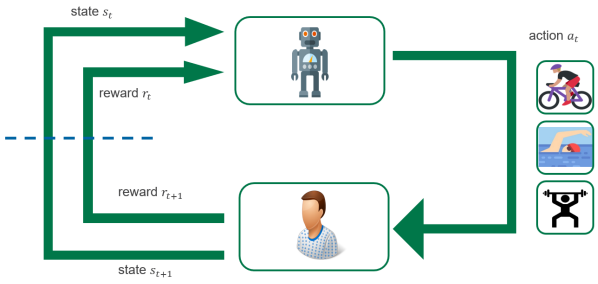

Figure 1: clinical pathway recommender

The process of creating a clinical pathway tailored to an individual patient spans several stages. To adapt a clinical pathway to a patient's needs, one starts from a disease specific blueprint and later incorporates the patients clinical picture as well as his or her individual preferences.
On an abstract level, the adaptation of a pathway can be modelled as a decision process. A number of activities must be decided upon, which in turn have interdependent effects among one another. Feedback on the effectiveness of the decisions made is often only given with a delay or in aggregated form - for example during a control visit to the doctor after a certain time. Reinforcement learning is about optimizing processes that can be described as a feedback control loop. The application of RL to the individualization of clinical pathways is therefore particularly well suited and promising.

The **personalization of a clinical pathway** is about identifying the optimal combination of activities and treatments in rehabilitation for an individual patient. In this context optimality can be considered from different viewpoints. On the one hand we see the fundamental objective of proposing rehabilitation measures that are safe from a medical perspective. On the other hand we aim to support the recovery process in the best-possible way by exploring alternative rehabilitation activities. While there are generic templates for different medical diagnosis - that are safe, there is the need to go beyond and adapt the clinical pathway - to provide tailored care plans to the individual patients.

In order to address the objectives described above for clinical path recommender systems, we present a **safety-aware reinforcement learning** approach. On a conceptual level this means that we have a state s_t of our patient and our agent proposes an action a_t - a rehabilitation measure - for our patient at time t (Figure 1).

The agent receives a reward r_{t+1} based on the change in the condition of our patient s_{t+1}. While classical RL is based on **try&error**, the healthcare application must **guarantee the safety of the patient** during the proposed activity. **Imitation Learning** is one of the ways in which this is pursued. Here the agent is trained to to „imitate" an expert's actions, i.e., to suggest a similar treatment activity to the one a doctor would choose faced with the same patient profile. Current work in imitation learning [11, 12] focus on efficiently learning from demonstrations while not paying special attention on safety or exploration. Research identified the objective safety in imitation learning[18] based their concept on being as close as possible to the examples shown. However, it is by no means guaranteed that the doctor's suggestion is optimal for the rehabilitation of the individual patient.

Challenges:

- How can we emphasize the importance of safety in suggesting rehabilitation treatments to a reinforcement learning agent?
- How can an agent explore the individual optimum and still remain within a safe and medically acceptable action space?

In answering the questions within this study we contribute to the following: **Contributions:**

- a conceptual approach to extract safety relevant behavior from expert demonstrations
- an adapted conceptual method for imitation learning that emphasizes safety-critical thinking
- implementation, application and preliminary evaluation of the concepts

Paper outline: After we position our work in the scientific related work in section 2, we introduce the conceptual background of our approach in section 3.1. While in section 3.2 we present the novel concepts of our approach, in section 3.3 we focus on the explicit application to optimize clinical pathways. In chapter 4 we outline our evaluation method and discuss the results achieved in chapter 5. Our work is then completed by a conclusion (section 6) and an outlook on future work in section 7.

2 RELATED WORK

Our work covers various areas of health care and machine learning, which we would like to examine in greater detail.

Research has shown an increasing interest in **applying machine learning techniques to health care** related tasks. From modelling disease progression [2] to automated clinical prognostics [1] methods of artificial intelligence have shown to be promising approaches. In further applications algorithms are used to annotate medical images and support doctors decision-making in a human-ML collaborative way [9]. Overall, decisive questions are emerging for the use of machine learning in the health sector. The decision of a system must be validated and made comprehensible. Only if the physician can be sure that the outcome of a machine learning algorithm is understandable and, above all, guarantees the safety of the patient, can systems prevail in the long term[16]. **Pathway - treatment** Bica et al. [6] introduced Counterfactual Recurrent Networks to estimate treatment effects by modelling treatment time-dependent impact on covariates based on the patient clinical history. Besides the topic-related relevant areas, various conceptual fields from machine learning are of relevance to our approach.

Imitation Learning is about training an agent to mimic the behaviour of an expert. With approaches such as inverse RL, e.g. GAIL - Generative Adversarial Imitation Learning - have recently achieved remarkable success [12]. Beyond this, we have seen approaches that attempt to reconstruct the expert's objective by evaluating hypotetic behaviour of an agent [22]. Further adaptations of imitation learning approaches are concerned with incorporating examples of an expert during the active learning process [4]. However these approaches neither have been adapted to learn from sub-optimal examples nor do they emphasise safety-relevant aspects.

constraint RL: First considerations about setting boundaries to the exploration of a reinforcement learning agent go back to the year 2000[3]. Recent work applied constraints in form of predefined threshold-values in continuous action spaces by adding a safety layer that in case of constraint violation corrects the suggestion of policy network [10]. Other then our approach, these concepts are based on pre-defined limits that are not deduced from examples and do not learn from experts.

safety RL: We have seen approaches that measure the similarity between the novice and the expert choice of action to prevent the agent from suggesting unsafe actions by considering the state distribution [19] or disagreement between multiple agents [7]. Follow up research did consider the quantification of policy uncertainty to model risk of exploration [20]. Lee et al. [15] proposed end-to-end imitation learning, where safety is addressed by evaluating the uncertainty of Bayesian convolutional network. Yet again, no approach has been adopted to differentiate existing demonstrations and adapt safety-relevant behaviour in a targeted manner.

multi criteria Laroche et al. [14] introduced a Multi-Advisor RL where n advisors are specialized on a sub task of the problem and an aggregator is used to derive a global policy based on the individual recommendations. While the safety of an RL problem can be described as a multi-criteria problem, the question remains to be answered how the approach described here can guarantee compliance with safety constraints and foster exploration within these limits.

3 OUR APPROACH

Contrary to previous imitation learning techniques, our approach focus on avoiding unsafe states while still exploring safe states to find the optimum. We teach the agent to handle safety critical states by imitating expert actions in similar situation. In safe states however the agent does not need to stick exactly with the behavior observed in expert demonstration. In fact we encourage it to search for the best personalized clinical path possible by exploration. While current safety RL algorithms [14, 19] focus on choosing actions that converge to the median of expert demonstrations that often is not the optimum, our approach aims at encouraging the agent to explore the state space while staying inside safe boundaries.

3.1 Formal Description

At each step t the agent selects an action $a_t \in A(s)$ based on the received representation of the environment *state* $s_t \in S$. Applied in a health recommender system speaking about action and states relates to recommendations for therapy activity and patients clinical state respectively. The agent receives a *reward,* r_{t+1}, which quantifies the development of the clinical condition and personal well-being of the patient and a new state s_{t+1} of the patient as a consequence of its action. $\pi_t(a|s)$ is the agents policy which is assigning a probability to each action at a given state and choices the most promising. This part is to be trained during the exploration or in the case of imitation learning during the expert observation. Since the new state serves as the input for the next iteration the agent keeps on interacting with the environment and creates a **trajectory** $\tau = (s_t, a_t)|t \in [t_0, t_h]$ where s_t is the state at a given time and a_t the action where t includes all elements from the start time t_0 to the time of termination t_h. The trajectory for an individual patient directly relates to the configured pathway (actions a_t map to the parameterised treatment activities foreseen in the clinical pathway) and the observed reaction of the patient (states s_t). The objective function denoted as $J(\pi)$ relates to

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}[R(\tau)] = R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t \tag{1}$$

where $\gamma \in [0, 1]$ is a discount factor. As we are dealing with the complex task of adapting clinical pathways, the modelling of several objectives and constraints gains in importance. **constraints** Constrained Markov Decision Processes (CMDPs) [3] limit the number of policies to a subset $\Pi_C \subset \Pi$ that fulfill a set of constraints C such that:

$$\Pi_C = \{\pi : J_{c_i}(\pi) \leq d_i \ \ \forall i = 1, \ldots, k\} \tag{2}$$

$$J_{c_i}(\pi)) = \mathbb{E}_{\tau\sim\pi}[c_i(\tau)] \tag{3}$$

J_{c_i} is the estimation of the expected value for a cost function c_i over the space of the trajectories achieved by the policy p_i. The resulting space of allowed policies is defined by the limitation that it only includes policies that do not exceed a defined limit $d_i \in \mathbb{R}$ for all the defined cost functions.

3.2 Safety Imitation

Focusing on modelling the safety of an reinforcement learning agent we define c_{safety} for brevity c_s to approximate the safety of a given state s_t. The flexibility of the approach provides the possibility to differentiate safety in several dimensions or to describe it as a holistic unit. In the case of Imitation Learning from sub-optimal but safe demonstrations we calculate the threshold value d_s over the distribution of expert trajectories, such that $d_s = max J_{c_s}(\pi_{exp})$ from the observed in the expert demonstrations T_{exp}. Evaluating the received expert trajectories we can now quantify how critical the different states were in terms of safety by defining:

$$T^{\epsilon}_{exp} = \{(s_t, a_t) : s_t \in T_{exp} \wedge J_{c_s}(s_t) \geq d_s - \epsilon\} \tag{4}$$

By focusing on the subset T^{ϵ}_{exp} to train our agent we can assure that it knows how to handle critical situations while preserving the freedom of exploring safe states. The collected demonstration data set is then weighted in such a way that the training data set for imitation learning consists of safety-relevant trajectories (T^{ϵ}_{exp}) to a defined extent mixed with randomly sampled trajectories from T_{exp}. Trough out this paper we will refer to this weighing as **safety focus** $\alpha \in [0, 1]$.

$$T^{train}_{exp} = \{\alpha * (s_t, a_t) \subset T^{\epsilon}_{exp} \cup (1-\alpha) * (s_t, a_t) \subset T_{exp} \setminus T^{\epsilon}_{exp}\} \tag{5}$$

It is essential to highlight that the data set used for the training of the agent is not extended by additional information such as the security factor, but rather a subset of the demonstrations is deliberately chosen for the training. The agent during imitation learning is not told at any time whether the state action pair currently presented to him in the context of supervised learning is a security relevant example. The approach changes solely the composition of the training data set.

3.3 Implication Health Recommender System

Our approach allows to learn effectively from demonstrations that guarantee a safe state of the environment respectively of the patient, but beyond that the actions not always show the optimal reaction to the state. Furthermore we aim to train a reinforcement learning agent with weighted expert demonstrations and thereby putting safety or other evaluation criteria in the foreground. Applying the approach on training a reinforcement learning agent to suggest and parameterize treatment activities in a clinical pathway we train the agent to explore the optimal recommendation while imitating expert recommendations when facing critical states described by constraint cost functions.

If the formal description is applied to the healthcare application, the cost function evaluates the clinical condition of the patient. In concrete terms, one could, for example, evaluate the deviation of the measured pulse from rest or optimal pulse. One now look at the expert's demonstration, i.e. any number of pairs of the patient's condition and the proposed therapy measure, you can evaluate for each demonstration what the cost function is, i.e. the safety assessment of the patient's clinical condition. It is crucial that the costs are not per se included in the objective function but are used as restrictions. As a result, an increased heart rate is not interpreted as negative by our recommender, but we take care in the decision-making process that the safety of this attribute is within certain limits.

We are therefore aware that the heart rate drops out during a therapeutic measure, and that this is one of the undesirable effects. But we want to make sure that the proposals of our intelligent system are within the limits of the experts' opinions. So if we see in the trajectories that the safety costs are below a certain level, we want to make sure that our proposals do not exceed this limit. To learn this, the demonstrations where the patient's condition was particularly close to the observed limit are particularly relevant. In our approach we define a sub set of trajectories T^{ϵ}_{exp} that have a defined distance from the critical limit ϵ. From this sub set we know that it is particularly relevant to learn how to avert critical states. During the training process, our intelligent system should accordingly pay special attention to adapting the expert suggestions close to the critical states.

4 EVALUATION

Although the concept presented here was developed out of the motivation to individualize clinical pathways for patients, it can be applied to various applications of reinforcement learning. For this reason, and because clinical data was not available to the extent necessary for an analysis, the evaluation is based on common and comparable safety problems in the directive. We use the gym environment provided by OpenAI.

4.1 Gym Environment

The gym environment offers the possibility to run different task scenarios for reinforcement agent and to extend the provided framework. Especially atari games and two dimensional games such as car racing are very popular and provide an excellent baseline to compare results. Due to the parallel use of the car racing environment as a recommender in the health care sector, the car racing environment is particularly suitable to demonstrate the functionality of our approach. The car on the race track describes the condition of the patient, who changes depending on the action - steering and accelerating, or parameterization of the next treatment measure. The more critically the condition of the patient - the position of the vehicle on the track - is evaluated, the more relevant it is to behave similar to the expert demonstrations. While we have described the relevance of heart rate in the clinical environment above, safety in this environment can be quantified with a cost function based on the distance to the edge of the track. So while in the medical case we can observe how a doctor behaves when the heart rate is particularly high or exceptionally low, in this environment we can quantify how far the vehicle is from the edge of the track.

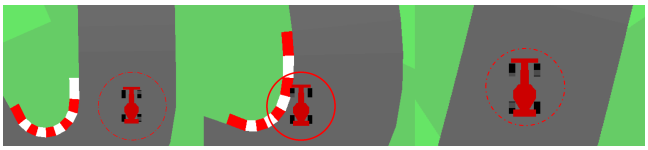

Figure 2: safety critical and uncritical states in the evaluation environment

To learn how to deal with critical conditions, we then look at the demonstrations where the assessment of the condition was particularly critical, as described in the formal description. The subset used for imitation learning is selected upon those based on equation 5.

4.2 Experiment Set Up

In the following we will describe the dimensions and parameters used for our evaluation in more detail: **Demonstrations**:All experiments were carried out on the same demonstration data set of size $|T_{exp}| = 4692 * (s_t, a_t)$, for further detail see Appendix A.
Imitation learning was performed as supervised learning of a tensorflow model with same architecture for every experiment. The agent was trained for 2000 *batches*, pairs of (s_t, a_t) respectively.
Cost function: In our evaluation we consider three **cost function** that quantify the cars position in the environment. Since the cars state represents a patients clinical state this can be seen as three different clinical parameters that are monitored during expert training. The cost functions considered quantify the cars position by evaluating the game frame received as a state representation. In the three directions $left$, $front$ and $right$ we calculate the distance to the unsafe state - the green besides the road. Evaluating these three cost functions for each state observed during the demonstration we develop a representation of the states safety. The parameter ϵ, which indicates how early a state should be classified as safety-relevant is set to $\epsilon = 5$. To calculate we move from the edge of the distribution of expert examples in the dimension of a constraint - in this case the safety - to the centre of the distribution. Visually this parameter defines how wide the edge of the distribution is, which is classified as safety critical as shown in Figure 3.

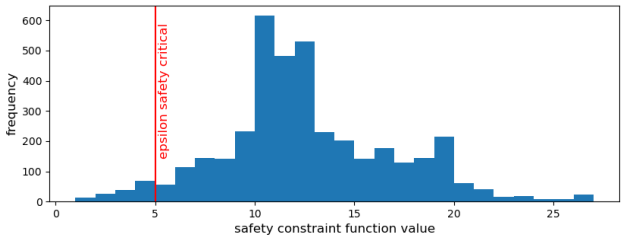

Figure 3: distribution safety demonstration

Agent testing After the weights of the reinforcement learning agent were trained via imitation learning the agent is evaluated in a newly generated gym environment. Here we observe the agent for two whole episodes to collect information about its performance and its safety. Depending on the individual performance of the agent this relates to ≈ 2000 state action pairs.

5 PRELIMINARY RESULTS

In the following we want to present the preliminary results of applying our approach to the safety critical decision process described in 4.2. Different values for α in equation 5 has shown significant influence on the performance of the agent with respect to the safety as well as the reward as shown in table 1.

Table 1: preliminary results safety focus

safety focus α	safety mean	safety std	reward mean
0.0	13.05	16.41	139.50
0.1	17.30	12.87	228.88
0.5	**21.37**	**11.12**	**697.41**
0.8	20.94	14.08	549.51

The results show that the safety focus has a significant impact on the agent's performance. The agent trained with the unweighted expert demonstrations achieves an average safety rating of 13.05 for its proposals, and the variation in safety over the ≈ 2000 state action pairs of 16.41 should be noted. The approach of pre-selecting and weighting the demonstrations based on the distribution of the cost function shows a positive impact. The security evaluation of the conditions caused by the agent can be raised to a level of 17.3 by

a weighting of $\alpha = 0.1$, and by a weighting of $\alpha = 0.5$ it can achieve a value of 21.37. In addition, the weighting of the expert trajectories in these cases also leads to a more robust reinforcement learning agent, which is reflected in the standard deviation of safety.

To make the results presented more comprehensible, Figure 4 provides a visualization.Training the agent with different **safety focuses** α results in safety and reward shown on the y-axis and the standard deviation represented by the dots size.

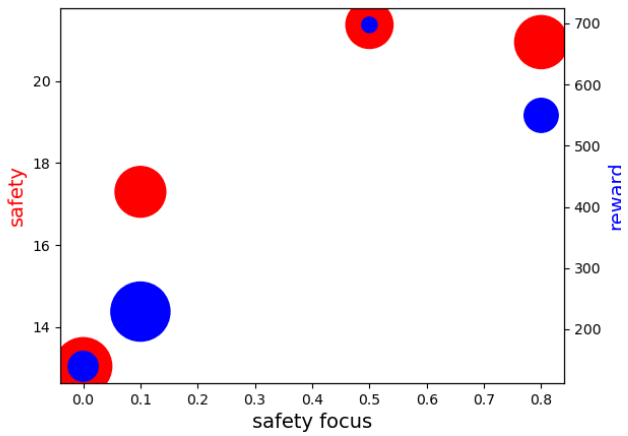

Figure 4: impact safety focus on episode reward and safety

Taking a closer look at comparing the cost function values for two agents - one trained **without safety focus** (5) and one trained with a **safety focus** $\alpha = 0.8$ (6) - emphasising the safety critical trajectories T^{ϵ}_{exp} in the expert demonstrations can significantly raise the safety of the actions recommended by the agent. While the performance of the agents is already reflected in the values listed in table 1, the reasons for this can be identified in Figures 5 and 6.

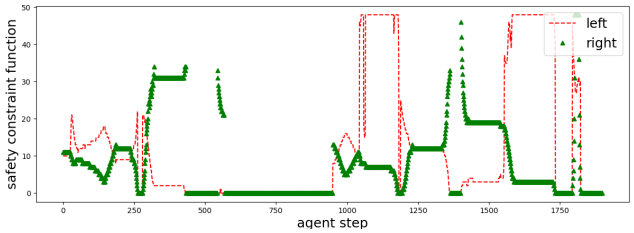

Figure 5: no safety focus

The non safety focus runs where not able to obtain a critical distance to the critical states,, while the safety focus runs successfully learned to avert critical states in an expert reaction manner.

While the agent without safety focus was not able to learn the correct handling of safety critical conditions during imitation learning, our approach was successful in adapting the expert's handling of critical states. By pre-selecting the expert examples without providing any further information during the training process, the agent with safety focus was able to avert safety critical conditions similar to the expert's behaviour.

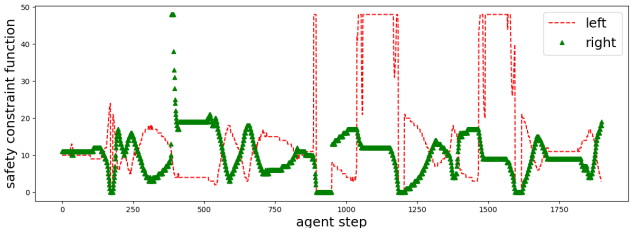

Figure 6: safety function 0.8 safety focus

6 CONCLUSION

The motivation for this work is derived from the medical context, in which the objective is to adapt clinical pathways to a patient's needs in the best possible way. while this scenario can be aptly described as a reinforcement learning problem, as discussed in the introduction, it is important to limit the exploration and thus the parameterisation of therapies and activities to a safe range of action from a medical point of view. the imitation learning approach offers a suitable approach to imitate the behaviour of experts. However, two central questions have arisen in reinforcement learning. Firstly, the question arose as to how an agent imitating an expert can concentrate on learning safety relevant actions. Furthermore, we asked ourselves whether an agent can be given the opportunity to explore the optimum within the action space while still maintaining a focus on safety.

To answer these questions, we have developed an approach that learns from expert demonstrations and concentrates on adapting the safety-relevant behaviour of the expert by appropriately weighting the examples provided. Our approach defines two parameters that determine how to deal with the state action pairs observed among experts. On the one hand, we have parameter ϵ, which indicates how early a state should be classified as safety-relevant. On the other hand we have **safety focus** α forcing the agent to train on a subset of expert trajectories, where α of the examples are classified as safety relevant under a given value ϵ.

Our approach for imitation learning was able to outperform equivalent agents trained on balanced demonstrations with regard to the safety as well as the reward. The generic conceptual approach underlying the work can be applied to a wide range of RM tasks. It is especially relevant for domains where expert knowledge is available, which defines how one should behave to be safe, but where it is not sure exactly what the optimal behaviour may look like. This is the case in the personalization of clinical pathways. while physicians can precisely advise which activities to suggest as rehabilitation under certain clinical conditions of the patient, it is not certain whether these suggestions are the optimal choice. with our approach we provide an important basis for exploring the optimum when proposing individually parameterized activities without violating the limits of the safety-relevant parameters.

7 FUTURE WORK

Besides the further exploration of the parameter combinations of ϵ and α, the transfer to additional RL problems is pending. Evaluating the approach on further 2D games in the gym environment is a logical next step. Additionally, teaching robotics to safely interact with their environment is relevant application [8]. Moreover, the approach is to be evaluated in more complex RL tasks that focus on the safety aspect, for which the recently published safety gym is available [21].

Future research should also consider how to completely avoid safety-critical examples that are dealt with by experts. One possible approach to this could be the simulation of responsibilities and the evaluation of possible reactions by an expert, using human in loop approaches as feedback for the system, see [17] and [5].

ACKNOWLEDGMENTS

This work was partially supported by the project vCare: Virtual Coaching Activities for Rehabilitation in Elderly (funded by Horizon 2020 research and innovation programme under Grant Agreement Number: 769807). Special acknowledgements are directed to the partners of the project, who have contributed valuable feedback in the specification of the research problem and by providing their expertise to this study.

REFERENCES

[1] Ahmed M. Alaa and Mihaela van der Schaar. 2018. AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning. *arXiv e-prints*, Article arXiv:1802.07207 (Feb. 2018), arXiv:1802.07207 pages. arXiv:1802.07207 [cs.LG]

[2] Ahmed M. Alaa and Mihaela van der Schaar. 2019. Attentive State-Space Modeling of Disease Progression. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 11338–11348. http://papers.nips.cc/paper/9311-attentive-state-space-modeling-of-disease-progression.pdf

[3] E. Altman. 1999. *Constrained Markov Decision Processes*. Chapman and Hall. https://doi.org/10.1016/0167-6377(96)00003-X

[4] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight Experience Replay. *arXiv e-prints*, Article arXiv:1707.01495 (July 2017), arXiv:1707.01495 pages. arXiv:1707.01495 [cs.LG]

[5] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L. Littman. 2019. Deep Reinforcement Learning from Policy-Dependent Human Feedback. *arXiv e-prints*, Article arXiv:1902.04257 (Feb. 2019), arXiv:1902.04257 pages. arXiv:1902.04257 [cs.LG]

[6] Ioana Bica, Ahmed M. Alaa, J. Brian Jordon, and Mihaela van der Schaar. 2020. Estimating Counterfactual Treatment Outcomes over Time Through Adversarially Balanced Representations. *In Proc. 8th International Conference on Learning Representations (ICLR 2020)* abs/2002.04083 (2020).

[7] Kiante Brantley, Wen Sun, and Mikael Henaff. 2020. Disagreement-Regularized Imitation Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkgbYyHtwB

[8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. arXiv:arXiv:1606.01540

[9] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. *arXiv e-prints*, Article arXiv:1902.02960 (Feb. 2019), arXiv:1902.02960 pages. arXiv:1902.02960 [cs.HC]

[10] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerík, Todd Hester, Cosmin Paduraru, and Yuval Tassa. 2018. Safe Exploration in Continuous Action Spaces. *CoRR* abs/1801.08757 (2018). arXiv:1801.08757 http://arxiv.org/abs/1801.08757

[11] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. 2017. One-Shot Visual Imitation Learning via Meta-Learning. *CoRR* abs/1709.04905 (2017). arXiv:1709.04905 http://arxiv.org/abs/1709.04905

[12] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 4565–4573. http://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning.pdf

[13] James Erica Snow Pamela Willis Jon Kinsman Leigh, Rotter Thomas. 2010. What is a clinical pathway? Development of a definition to inform the debate. *BMC Medicine* (2010).

[14] Romain Laroche, Mehdi Fatemi, Joshua Romoff, and Harm van Seijen. 2017. Multi-Advisor Reinforcement Learning. *CoRR* abs/1704.00756 (2017). arXiv:1704.00756 http://arxiv.org/abs/1704.00756

[15] Keuntaek Lee, Kamil Saigol, and Evangelos A. Theodorou. 2018. Safe end-to-end imitation learning for model predictive control. *CoRR* abs/1803.10231 (2018). arXiv:1803.10231 http://arxiv.org/abs/1803.10231

[16] Zachary C. Lipton. 2017. The Doctor Just Won't Accept That! *arXiv e-prints*, Article arXiv:1711.08037 (Nov. 2017), arXiv:1711.08037 pages. arXiv:1711.08037 [stat.ML]

[17] James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive Learning from Policy-Dependent Human Feedback. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 2285–2294. http://proceedings.mlr.press/v70/macglashan17a.html

[18] Kunal Menda, Katherine Rose Driggs-Campbell, and Mykel J. Kochenderfer. 2017. DropoutDAgger: A Bayesian Approach to Safe Imitation Learning. *ArXiv* abs/1709.06166 (2017).

[19] Kunal Menda, Katherine Rose Driggs-Campbell, and Mykel J. Kochenderfer. 2017. DropoutDAgger: A Bayesian Approach to Safe Imitation Learning. *CoRR* abs/1709.06166 (2017). arXiv:1709.06166 http://arxiv.org/abs/1709.06166

[20] Kunal Menda, Katherine Rose Driggs-Campbell, and Mykel J. Kochenderfer. 2018. EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning. *CoRR* abs/1807.08364 (2018). arXiv:1807.08364 http://arxiv.org/abs/1807.08364

[21] Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking Safe Exploration in Deep Reinforcement Learning. (2019).

[22] Siddharth Reddy, Anca D. Dragan, Sergey Levine, Shane Legg, and Jan Leike. 2019. Learning Human Objectives by Evaluating Hypothetical Behavior. *arXiv e-prints*, Article arXiv:1912.05652 (Dec. 2019), arXiv:1912.05652 pages. arXiv:1912.05652 [cs.CY]

A INSIGHT ON EXPERT DEMONSTRATIONS

Following we show the cost function calculated for the expert demonstrations. In 7 we see the to cost functions calculating the safety for $left$ and $right$.

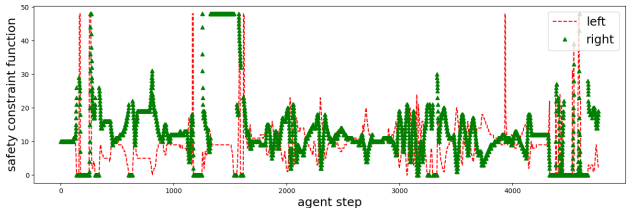

Figure 7: demonstration safety function left and right

In addition we evaluated the safety cost function in the dimension $straight$, as shown in 8.

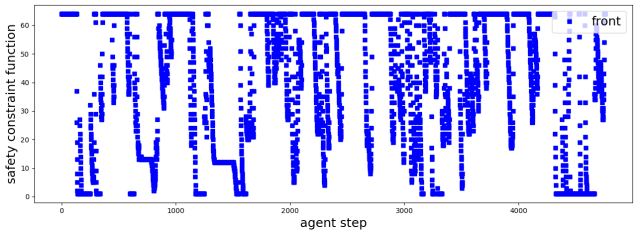

Figure 8: demonstration safety function straight

B ABLATION STUDY

In the following we provide further insights on the agents performance trained on **different levels of** $safetyfocus\alpha$.

Safety Focus 0.0 To complete the report on reinforcement agent performance with no safety focus besides 5 we provide the cost function referring to the safety evaluation $front$. Training the agent with $\alpha = 0.0$ results in the cost function to the front shown in Figure 9.

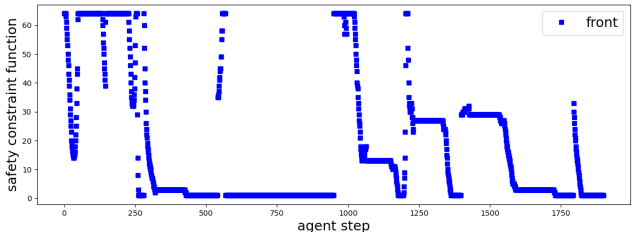

Figure 9: safety function front no safety focus

Safety Focus 0.1
Training the agent with a **safety focus of 0.1** results in the cost function shown below. Safety estimation to cost function sides is shown in Figure 10 and function front in Figure 11 respectively.

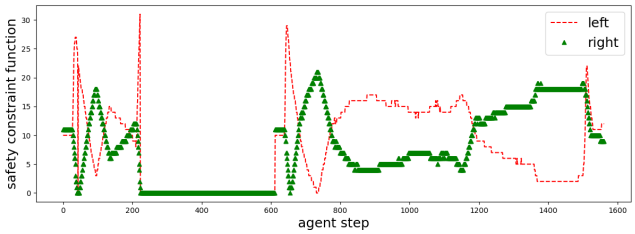

Figure 10: safety function side 0.1 safety focus

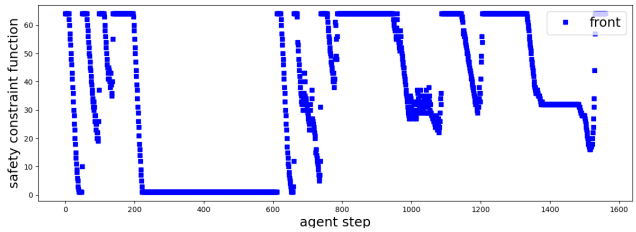

Figure 11: safety function front 0.1 safety focus

Safety Focus 0.5
Training the agent with a safety focus of 0.5 results in the safety function shown in Figure 12 for side safety estimation and 13 for c_{front} safety.

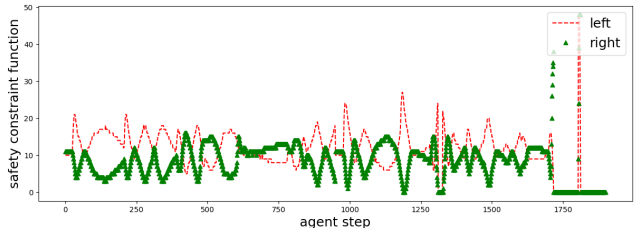

Figure 12: safety function side 0.5 safety focus

A Safety focus of 0.5 not only emphasises behavior to return from safety critical states with respect to the $left$ and $right$ safety constraint but also the $front$ safety.

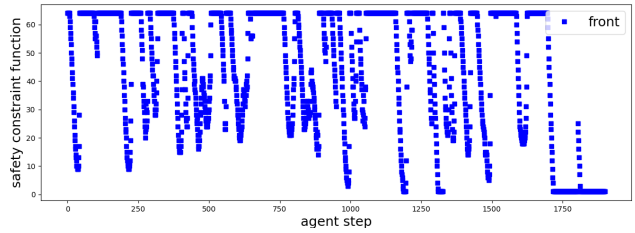

Figure 13: safety function front 0.5 safety focus

Safety Focus 0.8

In addition to the safety function values for $left$ and $right$ shown in 6 we provide the cost function for $front$. Training the agent with a safety focus of $\alpha = 0.8$ results in the cost function to the front shown in Figure 14.

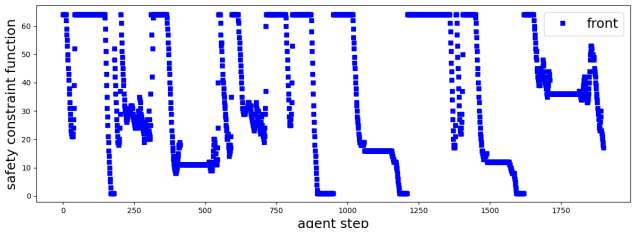

Figure 14: cost function front 0.8 safety focus