How to Train Your Agent: Active Learning from Human Preferences and Justifications in Safety-Critical Environments

Extended Abstract

Ilias Kazantzidis[1], Timothy J. Norman[1], Yali Du[2], Christopher T. Freeman[1]

[1]University of Southampton, Southampton, United Kingdom

[2]King's College London, London, United Kingdom

{ik3n19,t.j.norman,ctf1}@soton.ac.uk,yali.du@kcl.ac.uk

ABSTRACT

Training reinforcement learning agents in real-world environments is costly, particularly for safety-critical applications. Human input can enable an agent to learn a good policy while avoiding unsafe actions, but at the cost of bothering the human with repeated queries. We present a model for safe learning in safety-critical environments from human input that minimises bother cost. Our model, *JPAL-HA*, proposes an efficient mechanism to harness human preferences and justifications to significantly improve safety during the learning process without increasing the number of interactions with a user. We show this with both simulation and human experiments.[1]

KEYWORDS

Safe Reinforcement Learning; Learning from Human Preferences; Human-Agent Collaboration; Human-Robot Interaction

ACM Reference Format:

Ilias Kazantzidis[1], Timothy J. Norman[1], Yali Du[2], Christopher T. Freeman[1]. 2022. How to Train Your Agent: Active Learning from Human Preferences and Justifications in Safety-Critical Environments : Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

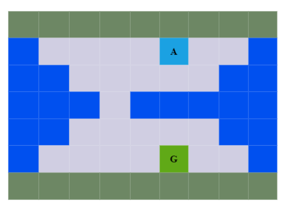
1 INTRODUCTION

The trial-and-error approach to training traditional Reinforcement Learning (RL) models, as used in complex games [15, 20, 22], is not suited for use directly in safety-critical environments. An agent exploring without taking into account safety can lead to damage to itself or its environment (e.g. in self-driving cars or robotics).

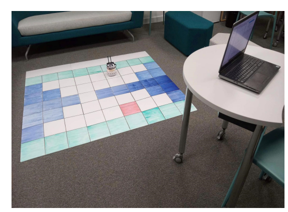
Safe RL addresses this Safe Exploration problem [9], and numerous techniques have been proposed that use information such as safety constraints [2, 4, 5, 10, 21]. These methods, however, do not offer both performance and safety guarantees *during training*. Our solution lies in the human-in-the-loop class and more specifically makes use of the technique where the agent learns from Human Preference queries. The agent samples two actions from its policy and asks the human which one they prefer. This is essentially an Active Learning [19] method and Christiano et al. [6] show why it is more sample efficient than other human-in-the-loop techniques such as Imitation Learning [13], Inverse Learning [1, 16], Reward Shaping [3, 23] and Human Intervention [18]. The main difference between our model and [6] is that the policy is learnt directly by supervised learning from an increasing dataset of human preferences, omitting the creation of a reward model learnt from human preferences and used with the traditional RL methods. Such a model would not tackle safe exploration as during optimisation the agent would continuously try unsafe actions to maximise the total reward.

Our contributions lie in two novel and generalisable ideas: (i) we augment preferences expressed by a human over a choice of actions with *justifications* such as one action is preferred because the other is unsafe; and (ii) we use these justifications to guide the generation of future queries over *hypothetical actions* (inspired by [17]), enabling the agent to more effectively map out unsafe areas.

As a learning paradigm and for the evaluation we use a modified version of the Island Navigation environment from [14] (initial state shown in Figure 1a) which, besides its small state-space, captures the safety semantics of the Safe Exploration problem well. Only horizontal and vertical actions are permitted, an episode ends with a death when the agent steps into a blue cell, and remains at the same place when it moves to a green cell. The goal is to avoid as many deaths as possible during training until the optimal policy (one that gives the fastest route to the goal state) is found.



(a) A: agent, G: goal, Dark blue cells: water, Green cells: wall, Grey cells: free road



(b) Human experiments with Thymio mobile robot (same configuration as on the left)

Figure 1: Island Navigation environment

2 JPAL-HA

Our algorithm, called ***J**ustified Human **P**references for **A**ctive **L**earning with **H**ypothetical **A**ctions* (***JPAL-HA***) (Algorithm 1) builds on the *Parenting* algorithm [8] with the main similarities being the direct policy learning from human preferences (Line 19) and the *parenting* query decision idea, i.e. 'the more familiar the agent's current state s_t is, the less likely it is to query from there, but instead act greedily' (Lines 11-13, 18, where $f(s_t)$ is the number of queries been issued from s_t). Our method uses: (i) *Before-The-Fact-Queries (BTFQs)*, i.e. queries issued to the human by the agent with dynamic probability before an action is taken (Lines 1-9); and (ii) *After-The-Fact Queries*

[1]The code can be found at https://github.com/ilkaza/JPAL-HA

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

Table 1: Average number of training deaths, etc. in simulation experiments until the optimal policy is found (Conservative setting, 1000 trials, mean and standard deviation $\mu \pm \sigma$)

	Parenting	*JPAL*	*JPAL-HA*
Training Deaths	0.07 ± 0.27	0.04 ± 0.23	0.02 ± 0.15
BTFQs	26.25 ± 5.8	31.12 ± 8.21	25.92 ± 6.53
Recordings	1.59 ± 1.45	1.97 ± 1.72	1.76 ± 1.64
ATFQs	1.32 ± 1.32	1.63 ± 1.54	1.47 ± 1.45
Overall steps	37.72 ± 14.25	41.94 ± 15.58	37.76 ± 14.5

Algorithm 1 JPAL-HA

Input: $p_{\text{BTFQ}} \in [0, 1)$: hyperparameter close to 1, p_{REC}: probability of recording, p_{ATFQ}: probability of asking an ATFQ
Output: $\pi(a|s)$: agent's policy model

1: **function** ask_BTFQ(s_t)
2: sample $a_t^{(0)}$ and $a_t^{(1)}$ and receive P and J_P from human
3: **if** $J_P = n$ **then**
4: find a^* and μ from P and J_P
5: add entry (s_t, $a_t^{(0)}$, $a_t^{(1)}$, μ, J_P, a^*) to X
6: execute a^*
7: **else if** $J_P = w$ **then**
8: $a_{best} \leftarrow$ gen_hypoth_actions(s_t, $a_t^{(0)}$, $a_t^{(1)}$, P, J_P)
9: execute a_{best}
10: **repeat**
11: **if** $(p_{\text{BTFQ}})^{f(s_t)} > r \overset{\text{iid}}{\sim} U[0, 1]$ **then**
12: ask_BTFQ(s_t)
13: **else**
14: **if** $p_{\text{REC}} > r \overset{\text{iid}}{\sim} U[0, 1]$ **then**
15: record a <*greedy, random*> action pair from s_t in R
16: **if** $p_{\text{ATFQ}} > r \overset{\text{iid}}{\sim} U[0, 1]$ **then**
17: ask an ATFQ from R and add entry to X
18: execute greedy action drawn from $\pi(a|s)$
19: train policy $\pi(a|s)$ with gradient descent minimising:

$$\mathcal{L} = -\sum_{s,a \in X} \sum_{i=0,1} \mu^{(i)} \log \frac{\pi(a_t^{(i)} | s_t)}{\pi\left(a_t^{(0)} | s_t\right) + \pi(a_t^{(1)} | s_t)} \tag{1}$$

20: **until** optimal policy is found

(ATFQs), i.e. queries *recorded* in a temporary memory R and issued to the human at a later step, with the goal to safely explore by eliciting the human's input upon the greedy and a random action (Lines 14-17). Answers of both types are stored in an embraced memory X, in which the agent policy model $\pi(a|s)$ (a neural network with the board state s as input and the probabilities of the four actions a as output) (Line 19) fits as in traditional supervised learning.

JPAL-HA introduces two novel ideas: (1) *Justifications*, $J_P \in \{w, n\}$, that augment preferences $P \in \{1^{\text{st}}, \textit{equal}, 2^{\text{nd}}\}$ over a choice of initially sampled actions $a_t^{(0)}$ and $a_t^{(1)}$ from $\pi(a|s)$. A warning, 'w', states that at least one action leads to a death, whereas 'n' indicates both are safe, making human feedback marginally more complex. The impact on safety, however, is significant. This is because a more efficient mapping of $P \times J_P$ to the correctly chosen action a^* and $\mu \in \{0, 0.25, 0.5, 0.75, 1\}$ (ground truth label with $\mu^{(0)} = \mu$ denoting how much the first action is preferred to the second and $\mu^{(1)} = 1-\mu$ denoting the opposite) is achieved compared to previous works which only use the values of $\mu \in \{0, 0.5, 1\}$; and (2) *Hypothetical Actions* (*HAs*): in case of J_P being 'w', then generate extra queries regarding new (hypothetical) sampled actions that could have been taken. By that, the agent investigates dangerous areas faster and potentially picks a better safe action a_{best} (Lines 7-9). A subtlety is that in most cases, only with a justification over the new action (i.e. 'if I had taken that other action, would it have been safe?') we can extract all the information we need in order to add a new entry in X about different actions involved in that step.

3 EVALUATION

Initially, we verified our assumption that traditional RL algorithms (e.g. Q-learning which suffers at least 20 deaths until it finds the optimal policy) fail in terms of safety. Table 1 shows the results of simulation experiments comparing *Parenting*, *JPAL* (*Justifications* only) and *JPAL-HA* (*Justifications* and *HAs*) on a *conservative setting* (high p_{BTFQ} value): $p_{\text{BTFQ}} = 0.95$, $p_{\text{REC}} = 0.8$ and $p_{\text{ATFQ}} = 0.8$. We observe that *Justifications* act immediately on safety (reduced training deaths), and *HAs* give a slight further improvement in safety with *Parenting* having higher training deaths than *JPAL-HA* with p-value $< 10^{-5}$ (*Dunn's test* [7] with *Holm-Bonferroni* adjustment [12]). Importantly, however, the use of *HAs* reduces human burden and time (BTFQs and overall steps) to a level not significantly different from *Parenting*. Additional experiments in a relaxed setting: $p_{\text{BTFQ}} = 0.8$, $p_{\text{REC}} = 0.5$ and $p_{\text{ATFQ}} = 0.5$ showed the same trend. For JPAL-HA we found 0.13 ± 0.35 training deaths (admissible increase) and 24.02 ± 5.98 BTFQs (decrease), revealing the trade-off between safety and human burden, which we can control by tuning p_{BTFQ} according to the application safety requirements.

We also conducted real-world training using the same configuration (Figure 1b) with 8 participants communicating via keyboard with a mobile robot [11]. We used the conservative JPAL-HA model and a seed with average values from Table 1. The mean time to complete training was $5\textit{min plus } 4s \pm 51s$ which we consider reasonable for training a real mobile robot. Moreover, we noticed that the cognitive effort of participants answering preference and justification queries was minimal. Additionally, the participants ran the same experiment exclusively on the computer. The mean time was $3\textit{min plus } 45s \pm 46s$. Assuming that most people could, after some practice, reach the best time which was $2\textit{min plus } 40s$ and dividing this by 27, i.e. the number of total queries, gives a good response rate of $5.9s/\textit{query}$ indicating the practicality of the method.

4 CONCLUSION

The novel ideas of Justifications and Hypothetical Actions, combined together in JPAL-HA have led to a significant improvement in safety, minimising the bother cost. Further experiments with JPAL-HA are being planned including real-world scenarios and incorporation of generalisable techniques such as transfer learning.

ACKNOWLEDGMENTS

This work was supported by UKRI [EP/S024298/1].

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. 1.

[2] Joshua Achiam, David Held, and Tamar. 2017. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 22–31.

[3] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. 2018. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748* (2018).

[4] Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3387–3395.

[5] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. 2018. A lyapunov-based approach to safe reinforcement learning. In *Advances in neural information processing systems*. 8092–8101.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*. 4299–4307.

[7] Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics* 6, 3 (1964), 241–252.

[8] Christopher Frye and Ilya Feige. 2019. Parenting: Safe reinforcement learning from human input. *arXiv preprint arXiv:1902.06766* (2019).

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

[10] Luca Gasparini, Timothy J Norman, and Martin J Kollingbaum. 2018. Severity-sensitive norm-governed multi-agent planning. *Autonomous Agents and Multi-Agent Systems* 32, 1 (2018), 26–58.

[11] MOBOTS group of the Swiss Federal Institute of Technology in Lausanne (EPFL) and the Lausanne Arts School (ECAL). last accessed Oct 2021. Thymio mobile robot. URL: https://www.thymio.org/.

[12] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[13] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–35.

[14] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI safety gridworlds. *arXiv preprint arXiv:1711.09883* (2017).

[15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[16] Andrew Y Ng, Stuart J Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[17] Siddharth Reddy, Anca Dragan, Sergey Levine, Shane Legg, and Jan Leike. 2020. Learning human objectives by evaluating hypothetical behavior. In *International Conference on Machine Learning*. PMLR, 8020–8029.

[18] William Saunders, Girish Sastry, and Andreas Stuhlmueller. 2018. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2067–2069.

[19] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[20] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.

[21] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. 2020. Safe reinforcement learning via curriculum induction. *Advances in Neural Information Processing Systems* 33 (2020), 12151–12162.

[22] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[23] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.