Reinforcement Learning from Human Feedback for Cyber-Physical Systems: On the Potential of Self-Supervised Pretraining

Timo Kaufmann[0000−0001−5193−8574], Viktor Bengs[0000−0001−6988−6186], and Eyke Hüllermeier[0000−0002−9944−4108]

LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
{timo.kaufmann,viktor.bengs,eyke}@ifi.lmu.de

Abstract. In this paper, we advocate for the potential of reinforcement learning from human feedback (RLHF) with self-supervised pretraining to increase the viability of reinforcement learning (RL) for real-world tasks, especially in the context of cyber-physical systems (CPS). We identify potential benefits of self-supervised pretraining in terms of the query sample complexity, safety, robustness, reward exploration and transfer. We believe that exploiting these benefits, combined with the generally improving sample efficiency of RL, will likely enable RL and RLHF to play an increasing role in CPS in the future.

Keywords: Reinforcement Learning from Human Feedback · Preference-Based RL · Self-Supervised Pretraining.

1 Introduction

Reinforcement learning (RL) considers the setting of learning behavior from rewarded interaction with an environment. The reward function specifies the desired behavior while the environment specifies the task dynamics. This setting is well-suited for cyber-physical systems (CPS), where the system repeatedly interacts with an environment to achieve some goal. RL can be used in this setting to learn a controller for a cyber-physical system, i.e., a policy that can choose appropriate actions based on the system's inputs. Examples of RL for CPS include applications to smart grids [18], Heating, Ventilation and Air Conditioning (HVAC) [32], energy storage [31], autonomous driving [3], as well as legged robots [43,39] and robotic manipulation [36].

One of the main challenges of applying RL to any task is measuring the agent's task performance in a way that is suitable for use as a reward function (reward design). Many of the largest successes of RL, such as as reaching or even exceeding human performance in the game of Go [37] and many Atari games [25], have been in the domain of games which have goals that are well-defined and easy to evaluate.

This is not the case for most real-world tasks however. Goals are often vague, subjective and characterized by trade-offs. Misspecifying these objectives can lead to surprising behaviors as well as safety issues [2]. Knox et al. [13] studies the challenges of reward design for autonomous driving, where the objective is a mixture of objective factors such as time to destination, fuel consumption and safety as well as subjective factors such as passenger experience. The right balance of these components may depend on context, such as time of day or the passenger's mood. More generally, Dulac-Arnold et al. [8] identifies reward design as one of the key challenges of applying RL to the real world.

RLHF is one way to cope with the challenge of reward design. Instead of assuming that a reward function is part of the problem specification, RLHF treats the reward function as part of the problem itself and attempts to learn it from human feedback. This is commonly done by collecting pairwise preference feedback over alternative agent trajectories (preference-based reinforcement learning (PbRL) [42]) and using it to infer a reward function, but other feedback modalities such as (imperfect) demonstrations [11], corrections [20], critiques [7] or natural language [41] may be used as well.

Examples of RLHF include ChatGPT [28], an instance of a large language model fine-tuned with RLHF to follow instructions [29] in a dialogue context. Other examples from the language domain are summarization [40] and question answering [27]. Beyond text, RLHF has been used to guide image generation [12]. RLHF has also been used in games [6] as well as simulated continuous control tasks [6,17]. In the domain of CPS, existing applications of RLHF include robot-to-human object handover [14] and robotic manipulation [5,38].

RLHF can greatly reduce the challenge of reward design by enabling us to learn tasks that humans can judge, even if they are difficult to express in an engineered reward function. This avoids the need to explicitly specify all objectives or their trade-offs – those can be communicated by example instead. The reward model can be trained to estimate human preferences directly from the system's sensor inputs. If the sensor inputs convey sufficient information, the agent can even learn different trade-offs for different contexts. For example, an internal camera in an autonomous vehicle could be used to judge the mood of the passenger or detect the presence of a child and adapt the driving behavior accordingly.

2 The Potential of Pretraining

Learning rewards directly from sensor inputs presents us with a new challenge however, since these sensor inputs (especially when they are vision-based) are often high-dimensional. High-dimensional state- and action spaces are already a challenge for RL without human feedback [8]. In that setting the problem is often tackled by data augmentation [44], representation learning [34,15] or model-based RL [10].

The latter two approaches – representation learning and model-based RL – can be considered instances of self-supervised learning [22,16], a form of learning that tries to learn something about the structure of the input data from unlabeled examples. This can be achieved by generating labels from the input data itself, such as training models to predict hidden parts of the input data or to determine whether two data points are related (e.g., transformations of each other) or not. Self-supervised learning is commonly used to learn representations or to initialize networks which are then later fine-tuned to specific tasks. Since self-supervised learning does not require any explicit human labels, it is possible to train on large amounts of data. This has been an important driving factor behind recent successes in the domain of language models [4].

In model-based RL, the self-supervised objective is to predict the environment dynamics, i.e., predict the next state from the current state and a chosen action. The goal of state-representation learning is to learn a representation of the agent's state that makes downstream tasks, such as reward prediction or policy learning, easier. Consider the example of an agent tasked with controlling an autonomous car: While the raw state of an agent may consist of low-level sensor inputs such as the pixels captured by a camera, the learned representation should capture information that is immediately relevant to the driving task such as the car's position relative to other cars and pedestrians in a higher-level format. Such a representation can be learned from data that is already available, such as experiences of the environment dynamics [34], and can then enable more sample-efficient learning of the downstream task, such as reward prediction. See the overview by Lesort et al. [19] for a more detailed introduction to state representation learning.

In this paper, we want to highlight the potential of self-supervised pretraining in the form of state representation learning and world model learning to effectively learn behavior from human feedback. We expect pretraining can improve query sample complexity as well as the learning system's safety and robustness, allow for better exploration of the reward function and enable transfer of knowledge between tasks.

Query sample complexity: Starting with a good state representation has the potential to learn more accurate reward models while requiring fewer human labels. Such a representation can be learned in a self-supervised manner from unlabeled interactions with the environment [34] or as a side-effect of model-based RL [26,10]. The learned representation is often more compact than the original observation and may also integrate information over multiple time-steps. This can be particularly beneficial in environments with high-dimensional observations such as images captured by a camera.

Similar sample-complexity benefits have been observed in RL without human feedback [34,45], where learned state representations can often decrease the necessary amount of interaction with the environment or even enable the application of RL to domains in which it was previously not feasible.

Metcalf et al. [24] explores this idea for RLHF and observes that by encoding environment dynamics in the state representation, i.e., choosing the representation learning task in such a way that the representation of the next state can be predicted from the current one with a simple linear layer, results in a significant increase in sample efficiency.

In addition to explicit representation learning, sample efficiency could also be improved through data augmentation [30] as well as semi-supervised learning [30].

Safety: Instead of learning a state representation in isolation, it is also possible to learn a full model of the environment dynamics (world model). A world model provides the option of synthesizing queries, i.e., generating hypothetical behavior for feedback. This changes the active learning setting from (repeated) pool-based sampling to membership query synthesis [1]. Since these trajectories can be tailored to be informative about the human preferences, this can increase the sample efficiency of the preference learning process. In addition, synthesizing queries can increase the safety of the learning process since potentially dangerous behavior can be tested without actually performing it in the real world. Needless to say that this is particularly important when working with physical systems. Initial work has explored the potential of synthesized queries in an RLHF context [33,23].

Another safety benefit of model-based RL is that it allows us to deploy separate policies in reality and in "imagination". Imagination refers to training that uses only interactions with the learned world model, not with the real environment. While the imagination policy may be focused on exploration, the real world policy may be focused on conservative data gathering.

Robustness: Synthesizing hypothetical behavior for feedback cannot only improve the system's safety, but may also contribute to the robustness and generalization of the learned rewards. This is because the synthesized queries can explore edge-cases that would rarely be encountered in the pool of experiences. It is possible to actively optimize the queries to fill gaps in the agent's knowledge of the human preferences. The benefits of membership query synthesis over pool-based active learning are discussed by Elreedy et al. [9].

Reward exploration: Model-based RL can be used to improve the exploration behavior of RL agents by learning an exploration policy that leads the agent to novel states purely in imagination, which can then be deployed in the real environment for efficient exploration. This avoids the issue of retrospective novelty, where RL agents with intrinsic exploration bonuses optimize their policy to visit states which they previously found novel – which, by definition, they are not anymore once they are included in the training data.

This approach has successfully been applied for regular state-space exploration [35]. Since reward-space exploration can be similarly important as state-space exploration for RLHF [21], one might expect additional benefits by applying this principle to reward-space exploration as well.

Transfer: Yet another benefit of representation- and model-learning is the possibility of transferring knowledge between tasks. Since a world model or state representation that was learned for one task remains valid for any other task with the same dynamics, this knowledge can be transferred and reward models for new tasks can be learned faster. A similar effect for model-based RL without human feedback is discussed by Moerland et al. [26].

3 Discussion and Conclusion

Learning controllers for cyber-physical systems has the potential of enabling many new use cases with complex interactions and increased integration of multiple systems. This may be of use for many applications, such as robotics, smart buildings and autonomous vehicles.

While to date applications of RL to real-world systems are sparse, the increasing sample efficiency of RL combined with the increased applicability to many tasks thanks to RLHF may cause that to change in the near future. Improving the feedback-efficiency of RLHF with approaches such as the ones discussed in this paper is therefore a promising area of future research. We believe that self-supervised pretraining has many benefits to offer and could play a crucial part in opening up many new use cases for cyber-physical systems.

Acknowledgements This research project/publication was supported by LMU-excellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria.

References

1. Aggarwal, C.C., Kong, X., Gu, Q., Han, J., Yu, P.S.: Active learning: A survey. In: Data Classification: Algorithms and Applications. CRC Press (2014). https://doi.org/10.1201/b17320-23
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P.F., Schulman, J., Mané, D.: Concrete problems in AI safety. CoRR **abs/1606.06565** (2016), http://arxiv.org/abs/1606.06565
3. Bai, Z., gen Cai, B., Shangguan, W., Chai, L.: Deep reinforcement learning based high-level driving behavior decision-making model in heterogeneous traffic. 2019 Chinese Control Conference (CCC) (2019)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (2020)
5. Cabi, S., Colmenarejo, S.G., Novikov, A., Konyushova, K., Reed, S., Jeong, R., Zolna, K., Aytar, Y., Budden, D., Vecerik, M., Sushkov, O., Barker, D., Scholz, J., Denil, M., de Freitas, N., Wang, Z.: Scaling data-driven robotics with reward sketching and batch reinforcement learning. In: Proceedings of Robotics: Science and Systems (2020). https://doi.org/10.15607/RSS.2020.XVI.076
6. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems (2017)
7. Cui, Y., Niekum, S.: Active reward learning from critiques. In: 2018 IEEE International Conference on Robotics and Automation, ICRA (2018). https://doi.org/10.1109/ICRA.2018.8460854

8. Dulac-Arnold, G., Levine, N., Mankowitz, D.J., Li, J., Paduraru, C., Gowal, S., Hester, T.: Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. Machine Learning **110**(9) (2021). https://doi.org/10.1007/s10994-021-05961-4
9. Elreedy, D., Atiya, A.F., Shaheen, S.I.: A novel active learning regression framework for balancing the exploration-exploitation trade-off. Entropy **21**(7) (2019). https://doi.org/10.3390/e21070651
10. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.P.: Mastering diverse domains through world models. CoRR **abs/2301.04104** (2023). https://doi.org/10.48550/arXiv.2301.04104
11. Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in Atari. In: Advances in Neural Information Processing Systems (2018)
12. Kazemi, H., Taherkhani, F., Nasrabadi, N.M.: Preference-based image generation. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2020 (2020). https://doi.org/10.1109/WACV45572.2020.9093406
13. Knox, W.B., Allievi, A., Banzhaf, H., Schmitt, F., Stone, P.: Reward (mis)design for autonomous driving. CoRR **abs/2104.13906** (2021), https://arxiv.org/abs/2104.13906
14. Kupcsik, A.G., Hsu, D., Lee, W.S.: Learning dynamic robot-to-human object handover from human feedback. CoRR **abs/1603.06390** (2016), http://arxiv.org/abs/1603.06390
15. Laskin, M., Srinivas, A., Abbeel, P.: CURL: Contrastive unsupervised representations for reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. vol. 119. PMLR (2020)
16. LeCun, Y., Misra, I.: Self-supervised learning: The dark matter of intelligence. Meta AI (2021), https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/, accessed: 2023-01-26
17. Lee, K., Smith, L.M., Abbeel, P.: PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021. vol. 139. PMLR (2021)
18. Lei, L., Tan, Y., Dahlenburg, G., Xiang, W., Zheng, K.: Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids. IEEE Internet of Things Journal **8** (2020)
19. Lesort, T., Díaz-Rodríguez, N., Goudou, J.F., Filliat, D.: State representation learning for control: An overview. Neural Networks **108**, 379–392 (2018). https://doi.org/10.1016/j.neunet.2018.07.006
20. Li, M., Canberk, A., Losey, D.P., Sadigh, D.: Learning human objectives from sequences of physical corrections. In: IEEE International Conference on Robotics and Automation, ICRA (2021). https://doi.org/10.1109/ICRA48506.2021.9560829
21. Liang, X., Shu, K., Lee, K., Abbeel, P.: Reward uncertainty for exploration in preference-based reinforcement learning. In: International Conference on Learning Representations (2022)
22. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE Trans. Knowl. Data Eng. **35**(1) (2023). https://doi.org/10.1109/TKDE.2021.3090866
23. Liu, Y., Datta, G., Novoseller, E.R., Brown, D.S.: Efficient preference-based reinforcement learning using learned dynamics models. CoRR **abs/2301.04741** (2023). https://doi.org/10.48550/arXiv.2301.04741

24. Metcalf, K., Sarabia, M., Theobald, B.: Rewards encoding environment dynamics improves preference-based reinforcement learning. CoRR **abs/2211.06527** (2022). https://doi.org/10.48550/arXiv.2211.06527
25. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M.A., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**(7540) (2015). https://doi.org/10.1038/nature14236
26. Moerland, T.M., Broekens, J., Plaat, A., Jonker, C.M.: Model-based reinforcement learning: A survey. Found. Trends Mach. Learn. **16**(1), 1–118 (2023). https://doi.org/10.1561/2200000086
27. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., Schulman, J.: WebGPT: Browser-assisted question-answering with human feedback. CoRR **abs/2112.09332** (2021), https://arxiv.org/abs/2112.09332
28. OpenAI: ChatGPT: Optimizing language models for dialogue (2022), https://openai.com/blog/chatgpt/, accessed: 2023-01-23
29. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems (2022)
30. Park, J., Seo, Y., Shin, J., Lee, H., Abbeel, P., Lee, K.: SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In: International Conference on Learning Representations (2022)
31. Qi, B., Rashedi, M., Ardakanian, O.: EnergyBoost: Learning-based control of home batteries. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems, e-Energy 2019 (2019). https://doi.org/10.1145/3307772.3328279
32. Raman, N.S., Devraj, A.M., Barooah, P., Meyn, S.P.: Reinforcement learning for control of building HVAC systems. In: 2020 American Control Conference, ACC 2020. IEEE (2020). https://doi.org/10.23919/ACC45564.2020.9147629
33. Reddy, S., Dragan, A.D., Levine, S., Legg, S., Leike, J.: Learning human objectives by evaluating hypothetical behavior. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. vol. 119. PMLR (2020)
34. Schwarzer, M., Anand, A., Goel, R., Hjelm, R.D., Courville, A.C., Bachman, P.: Data-efficient reinforcement learning with self-predictive representations. In: International Conference on Learning Representations (2021)
35. Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., Pathak, D.: Planning to explore via self-supervised world models. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. vol. 119. PMLR (2020)
36. Sermanet, P., Xu, K., Levine, S.: Unsupervised perceptual rewards for imitation learning. In: International Conference on Learning Representations (2017)
37. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T.P., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529** (2016)

38. Singh, A., Yang, L., Finn, C., Levine, S.: End-to-end robotic reinforcement learning without reward engineering. In: Robotics: Science and Systems XV (2019). https://doi.org/10.15607/RSS.2019.XV.073
39. Smith, L.M., Kostrikov, I., Levine, S.: A walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. CoRR **abs/2208.07860** (2022). https://doi.org/10.48550/arXiv.2208.07860
40. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize from human feedback. CoRR **abs/2009.01325** (2020), https://arxiv.org/abs/2009.01325
41. Williams, E.C., Gopalan, N., Rhee, M., Tellex, S.: Learning to parse natural language to grounded reward functions with weak supervision. In: 2018 IEEE International Conference on Robotics and Automation, ICRA (2018). https://doi.org/10.1109/ICRA.2018.8460937
42. Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J.: A survey of preference-based reinforcement learning methods. J. Mach. Learn. Res. **18** (2017), http://jmlr.org/papers/v18/16-634.html
43. Wu, P., Escontrela, A., Hafner, D., Goldberg, K., Abbeel, P.: DayDreamer: World models for physical robot learning. CoRR **abs/2206.14176** (2022). https://doi.org/10.48550/arXiv.2206.14176
44. Yarats, D., Kostrikov, I., Fergus, R.: Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In: International Conference on Learning Representations (2021)
45. Yu, T., Lan, C., Zeng, W., Feng, M., Zhang, Z., Chen, Z.: PlayVirtual: Augmenting cycle-consistent virtual trajectories for reinforcement learning. In: Advances in Neural Information Processing Systems (2021)