The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^2HCAI)

A State Augmentation based approach to Reinforcement Learning from Human Preferences

Mudit Verma[1] **Subbarao Kambhampati** [1]

[1] SCAI, Arizona State University
muditverma@asu.edu, rao@asu.edu

Abstract

Reinforcement Learning has suffered from poor reward specification, and issues for reward hacking even in simple enough domains. Preference Based Reinforcement Learning attempts to solve the issue by utilizing binary feedbacks on queried trajectory pairs by a human in the loop indicating their preferences about the agent's behavior to learn a reward model. In this work, we present a state augmentation technique that allows the agent's reward model to be robust and follow an invariance consistency that significantly improved performance, i.e. the reward recovery and subsequent return computed using the learned policy over our baseline PEBBLE. We validate our method on three domains, Mountain Car, a locomotion task of Quadruped-Walk, and a robotic manipulation task of Sweep-Into, and find that using the proposed augmentation the agent not only benefits in the overall performance but does so, quite early in the agent's training phase.

Introduction

Machine Learning (Pouyanfar et al. 2018; Verma and Buduru 2020; Verma et al. 2019b), specifically Deep Reinforcement Learning (Mnih et al. 2015; Verma et al. 2019a; Arulkumaran et al. 2017) has been quite successful for several tasks (Brockman et al. 2016) that especially those with high dimensional states and action spaces. However, much of the credit to the RL agent's success is attributed to the reward function specification (Hadfield-Menell et al. 2017; Abel et al. 2021; Ferret et al. 2020; Verma et al. 2019a). Among several challenges in designing a "good reward" (Hadfield-Menell et al. 2017) function that allows the RL agent to learn trustworthy policies (Zahedi et al. 2021, 2022) like sparsity of the reward, balancing scales of reward signals, etc, one major issue that researchers have found is the inability of specifying the reward function, to begin with, (Vamplew et al. 2018). To allow humans in the loop to specify their preferences over the agent's goals and behavior, Preference-Based Reinforcement Learning (PbRL) (Wilson, Fern, and Tadepalli 2012) utilizes binary evaluative signals over trajectory pairs given by the human in the loop to convey their preferences to learn the human's reward model and subsequently learn a usable policy over it.

Preference Based Reinforcement Learning allows human in the loop to specify their preferences in the form of binary feedback over agent-generated trajectory pair queries. This can prove to be quite beneficial for several reasons. First, the human in the loop gets to realize how the trajectories "look" like and does not have to imagine these trajectories thereby taking into account human's cognitive limitations. Second, the human in the loop need not be an expert engineer and is not required to be aware of the underlying agent representation to hand-design a reward function and instead can use the image as the lingua franca (Guan et al. 2021; Guan, Verma, and Kambhampati 2020; Kambhampati et al. 2022) to specify their preferences. Finally, issues of reward hacking and "cheat-behaviors" have been noted by researchers in prior works and PbRL can serve as a potential alternative to reward engineering to tackle these issues (Vamplew et al. 2018; Krakovna et al. 2020).

Recent works in PbRL have focused on improving the query strategy (Lee, Smith, and Abbeel 2021; Christiano et al. 2017), exploration method (Lee, Smith, and Abbeel 2021), priors regarding the reward function (Verma and Metcalf 2022), semi-supervised learning methods (Park et al. 2022) and temporal data augmentation (Park et al. 2022). However, to the best of our knowledge, none of the existing methods have attempted to provide a tailored solution for the problem PbRL with image-based state representations. In this work, we specifically focus on the problem of learning from human preferences via binary feedback on trajectory pair queries with a pixel-based image state representation of the agent (Mnih et al. 2015, 2013). This is particularly helpful as in most of the recent PbRL works the lingua franca used to communicate with the human in the loop are image-based representations. Motivated by prior works in explainable AI (Verma, Kharkwal, and Kambhampati 2022) and Advisable reinforcement learning (Sreedharan et al. 2020; Guan et al. 2021; Guan, Verma, and Kambhampati 2020; Verma et al. 2021), we present a data augmentation technique over image-based state representation of the agent that shows significant improvements in the agent's reward recovery and subsequently high return collected by the agent's learned policy over the learned reward model for three continuous control tasks, Mountain-Car, a robotic manipulation task Sweep-Into, and a locomotion task Quadruped-Walk.

Data augmentation has been explored for various reasons in the context of machine learning and artificial intelligence (Shorten and Khoshgoftaar 2019; Srinivas, Laskin, and Abbeel 2020; Guan et al. 2021), of which the foremost reasons are the robustness of the learned model (Cobbe et al. 2019) and to impose consistencies (Sohn et al. 2020; Xie et al. 2020; Berthelot et al. 2019). In addition to the requirements of the robustness of the reward model and to impose certain invariance consistencies we leverage the observation that a good reward model should "focus" on regions of importance and this should be reflected by the model's representation space. We combine the requirements of robustness, invariance consistency and to have the reward model "focus" on regions of importance into a single data augmentation technique and showcase the benefits across three continuous control tasks (OpenAI Gym's Mountain Car (Brockman et al. 2016), DM Control's locomotion task of Quadruped-Walk (Tunyasuvunakool et al. 2020) and Metaworld's robotic manipulation task of Sweep-Into (Yu et al. 2020)) and show that using such an augmentation technique can help with an early boost in learning performance at zero cost to the human in the loop.

Background

We have an agent M that interacts with the environment $\mathcal{E}$ by taking an action $a \sim \mathcal{A}$ on a state s. Typical reinforcement learning frameworks assume an underlying Markov Decision Process (MDP) as the tuple $<\mathcal{S}, \mathcal{T}, \mathcal{A}, \mathcal{R}>$ where $\mathcal{S}$ is the state space, $\mathcal{T}$ is the transition function, $\mathcal{A}$ is the set of permissible actions and $\mathcal{R}$ is the task reward function, however Preference-based Reinforcement learning updates this tuple by replacing $\mathcal{R}$ with $\tilde{\mathcal{R}}^{\text{h}}$, the human reward model, as the tuple $<\mathcal{S}, \mathcal{T}, \mathcal{A}, \tilde{\mathcal{R}}^{\text{h}}>$.

The goal of PbRL is to approximate the human reward model $\tilde{\mathcal{R}}^{\text{h}}$ with a parameterized function approximator R^h_ψ. The agent, M, queries the human in the loop with a trajectory pair τ_0, τ_1, $\tau_i = \{(s_k, a_k), (s_{k+1}, a_{k+1} \cdots (s_{k+H}, a_{k+H}))\}$ and receives a binary feedback $y \in \{0, 1\}$ indicating the human's preferred trajectory, i.e. $y = 0$ if τ_0 is preferred over τ_1 and vice versa. Such feedbacks along with the queried trajectories are stored in a dataset D_τ as tuples (τ_0, τ_1, y). Recent PbRL works have leveraged the Bradley Terry model (Bradley and Terry 1952) to compute the probability of one trajectory being preferred over another. With means to computing this probability, PbRL methods treat essentially solve the reward learning problem via a classification problem where the trajectory with a higher approximated sum of rewards (or the return) is predicted to be the human preferred trajectory and a reward assignment to the constituent states of τ_0, τ_1 that achieves high accuracy in this supervised learning task is taken to be the approximate human reward model. As typical in binary classification problems of supervised learning, this is done by minimizing the cross-entropy between the predictions and ground truth human labels as follows:

$$\mathcal{L}_{CE} = \underset{(\tau_0,\tau_1,y)\sim\mathcal{D}}{-\mathbb{E}}[y(0)\log P_\psi[\tau_0 \succ \tau_1] + y(1)\log P_\psi[\tau_1 \succ \tau_0]] \quad (1)$$

where probabilities P_ψ are computed using the approximated reward function R^h_ψ as :

$$P_\psi[\tau_0 \succ \tau_1] = \frac{\exp\Big(\sum_t R^h_\psi(s^0_t, a^0_t)\Big)}{\sum_{i\in\{0,1\}} \exp\Big(\sum_t R^h_\psi(s^i_t, a^i_t)\Big)} \quad (2)$$

Method

The key contribution of this work is to allow the reward model to better utilize the queried trajectory pairs in the context of image-based state representation of the RL agent M. We aim to leverage three key benefits from a single state augmentation technique, namely, **robustness**: a key indicator for generalization of the learned rewards which becomes all the more necessary with the aim of reducing the human feedback sample complexity, **invariance**: to allow the agent to successfully learn the state representation in the reward model such that it is invariant to perturbations to regions in the image space which is not important, and finally, **motion-based-importance**: that marks regions in the image observations of the agent where a change has occurred intending to motivate the agent to "focus" on such potential regions while predicting the reward.

As noted by several prior works in the area of explainable AI and advisable RL, image observations are expressed as the pair $< I_C, I_S >$ where I_C are the pixels denoting the content of the image and I_S denotes the pixels that inform about the style of the image. Typically, pixels related to the style of the image do not offer any additional information than just I_C and I_C would conceptually be similar to the effective dimensionality of the image observation required for the task in concern. Prior literature has utilized either human annotations to assume I_C or attempted to use explanatory techniques to inform users about I_C. In this work, we use the "motion-based-importance" insight to capture I_C. For an image observation and agent state I, let's say $I_1, I_2 \cdots I_n$ are observations such that there exists some action a_i which when taken on I_j $j \in 1, 2 \cdots n$ can transition to I. Then the image mask created by the union (of boolean mask matrices) of all the differences between I_j and I will contribute to I_C, i.e.

$$\mathbb{M}(I, \{I_1, I_2 \cdots I_n\}) = \bigcup_{j\in\{1,2\cdots n\}} \mathbb{M}(I, I_j) \quad (3)$$

$$\mathbb{M}(I, I_j) = \begin{cases} 1, I(x,y) \neq I_j(x,y) \\ 0, \text{otherwise} \end{cases} \quad (4)$$

As the access to all predecessor states I_j for every I may entail high exploration and storage requirements, it can be approximated by using one predecessor observation I_{t-1} at a time to create the mask $\mathbb{M}$ for observation I_t in a trajectory $\tau =< I_0, I_1, I_2 \cdots I_n >$.

With equation 3 giving us the mask (as a proxy for all the pixels referring to the content I_C of an image I), we can utilize it to inject our requirements of invariance. A popular means of doing so is via perturbations of the style pixels. We follow, (Guan et al. 2021; Greydanus et al. 2018) which argues for using Gaussian perturbations as would still preserve the texture of the remaining images, however, domain-specific perturbations such as a change in objects in the background are also applicable but require extra information. Gaussian perturbations using the mask in equation 4 can be defined as :

$$\phi(I, I_j, \mathbb{M}) = I \odot (1 - \mathbb{M}(I, I_j) + \mathcal{G}(I, \sigma_{\mathcal{G}}) \odot \mathbb{M}(I, I_j) \tag{5}$$

In the equation 5, $\mathcal{G}$ refers to the convolution of a gaussian filter over the image I with standard deviation $\sigma_{\mathcal{G}}$. $\odot$ is the Hadamard product operation between the input image or the gaussian blurred image and the binary mask matrix $\mathbb{M}$. Finally, from the dataset of preference feedback $D_\tau =< \tau_0^i, \tau_1^i, y^i >$, we can pick each trajectory τ_0^i or τ_1^i and produce augmented trajectories by augmenting individual states and chaining those states together to form the new augmented trajectory. We augment a trajectory $\tau =< s_0, s_1, ...s_n >$ as,

$$\tau_g =< s_0, \phi_{\mathbb{M}}(s_1, s_0), \phi_{\mathbb{M}}(s_2, s_1) \cdots \phi_{\mathbb{M}}(s_n, s_{n-1}) > \tag{6}$$

and produce multiple augmented trajectories by varying the standard deviation $\sigma_{\mathcal{G}}$ of the gaussian filter. Hence for a set of standard deviations $\mathcal{G} = \{\sigma_1, \sigma_2 \cdots \sigma_{k-1}\}$ we can generate $k-1$ augmented trajectories from the original trajectory τ as, $\{\tau_g^1, \tau_g^2 \cdots \tau_g^{k-1}\}$ to get a total of k trajectories (one original τ, and $k-1$ augmented τ_g^i).

We impose our invariance consistency requirement via the following loss over the reward model R_ψ^h,

$$\mathcal{L}_I(\tau) = \frac{1}{|\mathcal{G}|} \sum_{i \sim \mathcal{G}} \left\| \mathbf{R}_\psi^h(\tau) - \mathbf{R}_\psi^h(\tau_g^i) \right\|_2 \tag{7}$$

where $\mathbf{R}_\psi^h(\tau) = \begin{bmatrix} R_\psi^h(s_0) & R_\psi^h(s_1) & \cdots & R_\psi^h(s_T) \end{bmatrix}^T$ is a vector of predicted rewards over the states in the trajectory. This reward $\mathcal{L}_I$ helps with both robustness and invariance consistency as shown in section . This final reward model update loss is a linear combination of this invariance loss and the cross entropy loss as,

$$\mathcal{L}_{reward} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_I \mathcal{L}_I \tag{8}$$

Finally, we also treat augmented trajectories as additional data points for training the cross entropy loss $\mathcal{L}_{CE}$. Since we optimize using stochastic gradient descent, we first sample a batch of trajectory pairs and labels from D_τ and then use augmented versions of these trajectories as additional data points within this batch.

Experiments

In the preliminary investigation, we were interested to answer whether the proposed augmentation technique can outperform the current state-of-the-art PbRL method (PEBBLE).

We validate our results on three continuous control domains on pixel-based state representations (with standard preprocessing (Mnih et al. 2015, 2013)), i.e. Mountain Car (`MountainCarContinuous-v0`) by OpenAI gym (Brockman et al. 2016), one locomotion task of Quadruped Walk `quadruped-walk` from MuJoCo (Tunyasuvunakool et al. 2020) and one robotic manipulation task Sweep Into `sweep-into-v2` from Metaworld (Yu et al. 2020). We use PEBBLE as the backbone Preference-based Reinforcement Learning algorithm and update the $\mathcal{L}_{reward} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_I \mathcal{L}_I$ with $\lambda_{CE} = 1, \lambda_I = 0.6$, and as mentioned before, use the augmented trajectories data for training the cross-entropy loss as well by appending to the dataset D_τ the following tuples $< \tau_g^0, \tau_g^1, y >_{i=1}^{i=|\mathcal{G}|}$ for each $< \tau_0, \tau_1, y >$. PEBBLE uses SAC as the Reinforcement Learning algorithm underneath, and we used the same hyperparameters used for Sweep Into and Quadruped-Walk as suggested in PEBBLE (although they primarily showcase results on low-level states instead of image observations). We gave 1000 feedbacks for Quadruped-Walk and Mountain Car, and 10000 feedbacks for Sweep-Into. SURF shows results on pixel-based inputs for PEBBLE achieve similar performance as our implementation of PEBBLE baseline. Finally, as suggested in prior works, we use an oracular approach to evaluation where the rewards that come packaged with these environments are assumed to be human's reward model, and the extent to which a PbRL method can recover this underlying reward model is used as the success criterion.

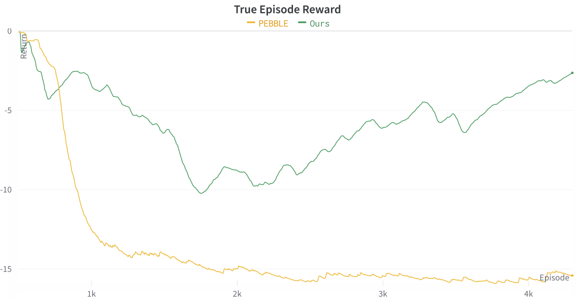

Figure 1: Evaluation curves on Mountain Car Continuous Control as measured on the ground truth human reward $\tilde{\mathcal{R}}^{\mathrm{h}}$.

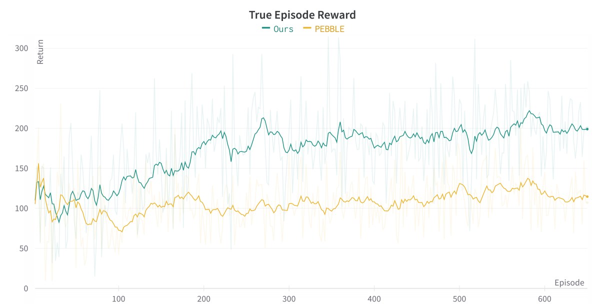

Figure 2: Evaluation curves on the locomotion task of Quadruped as measured on the ground truth human reward $\tilde{\mathcal{R}}^{\mathrm{h}}$.

Results

In order to realize how well a PbRL framework learned the reward model, we evaluate the policy π learned on R_ψ^h over the human's underlying reward model $\tilde{\mathcal{R}}^{\mathrm{h}}$. Any improvements over the performance of the learned policy on the underlying human reward model $\tilde{\mathcal{R}}^{\mathrm{h}}$ would indicate a more

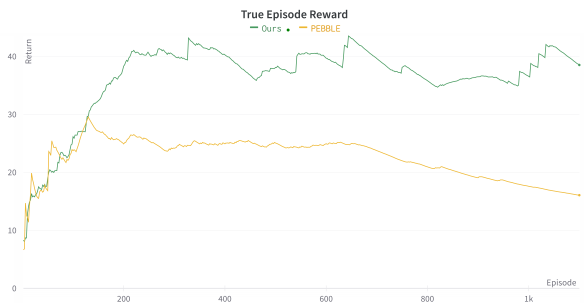

Figure 3: Evaluation curves on the robotic manipulation task of Sweep Into as measured on the ground truth human reward $\tilde{\mathcal{R}}^{\mathrm{h}}$.

meaningful learned reward model R^h_ψ.

Mountain Car: is used as a toy domain to verify our claims. The domain consists of a car placed at the bottom of a sinusoidal valley with the goal to strategically accelerate the car (action space is acceleration between -1 to 1) to reach the top of the right hill (Brockman et al. 2016). Figure 1. We find that with very few feedbacks taken in initial epochs, our approach can learn a reasonable reward model and with more human feedbacks it outperforms the baseline.

Quadruped-Walk: is a quadruped (two bipedals) robot, with four legs each having three actuators (a total of 12 continuous actions) with the goal of walking on a flat surface. Figure 2 shows that our approach outperforms PEBBLE baseline on this locomotion task and as seen before, the highest performance boost occurs within the initial training episodes.

Sweep-Into: is a robotic arm in front of a puck placed on a table with the goal to "sweep" the puck into the goal location. The puck positions are randomized at the start of each episode (as per the default package setup). We find similar results as in our other two domains, where the proposed state/trajectory augmentation (and subsequently the updated reward loss) outperforms the baseline PEBBLE. Additionally, we notice that the augmentations provide a significant boost to the performance early in the training process and help the agent maintain this performance gain throughout.

Discussion

In this work, we presented a state augmentation technique tailored for image-based Preference-based Reinforcement Learning. The proposed augmentation utilizes the insight that regions of the image observation that update upon a transition (when an action is taken), are at least a subset of all the "content" available in the image observation. This allows the reward model being learned to be cognizant of potential regions of the image observation that are likely to be updated in future steps (and history) while predicting the rewards. Augmented trajectories are treated as additional query data points to train the cross entropy loss along with the proposed invariance loss that maintains prediction consistency (Guan et al. 2021; Xie et al. 2020). The benefits of the approach were validated on three continuous control domains, OpenAI gym's Mountain Car, a locomotion domain: DM Control's Quadruped-Walk, and finally a robotic manipulation domain: Meta World's Sweep-Into.

Future work includes an exhaustive evaluation of the proposed augmentation across more domains. We also plan to evaluate the benefits of the proposed work when used with other PbRL techniques and paradigms. Finally, we would also like to study the advantages of using domain-dependent perturbations.

Acknowledgements

Kambhampati's research is supported by the J.P. Morgan Faculty Research Award, ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067 and DARPA SAIL-ON grant W911NF19-2-0006.

References

Abel, D.; Dabney, W.; Harutyunyan, A.; Ho, M. K.; Littman, M.; Precup, D.; and Singh, S. 2021. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34: 7799–7812.

Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cobbe, K.; Klimov, O.; Hesse, C.; Kim, T.; and Schulman, J. 2019. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, 1282–1289. PMLR.

Ferret, J.; Marinier, R.; Geist, M.; and Pietquin, O. 2020. Self-Attentive Credit Assignment for Transfer in Reinforcement Learning.

Greydanus, S.; Koul, A.; Dodge, J.; and Fern, A. 2018. Visualizing and understanding atari agents. In *International conference on machine learning*, 1792–1801. PMLR.

Guan, L.; Verma, M.; Guo, S. S.; Zhang, R.; and Kambhampati, S. 2021. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. *Advances in Neural Information Processing Systems*, 34: 21885–21897.

Guan, L.; Verma, M.; and Kambhampati, S. 2020. Explanation augmented feedback in human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2006.14804*.

Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S. J.; and Dragan, A. 2017. Inverse reward design. *Advances in neural information processing systems*, 30.

Kambhampati, S.; Sreedharan, S.; Verma, M.; Zha, Y.; and Guan, L. 2022. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12262–12267.

Krakovna, V.; Uesato, J.; Mikulik, V.; et al. 2020. Specification gaming: The flip side of AI ingenuity— DeepMind.

Lee, K.; Smith, L.; and Abbeel, P. 2021. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Park, J.; Seo, Y.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2022. SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning. *arXiv preprint arXiv:2203.10050*.

Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M. P.; Shyu, M.-L.; Chen, S.-C.; and Iyengar, S. S. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5): 1–36.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Sreedharan, S.; Soni, U.; Verma, M.; Srivastava, S.; and Kambhampati, S. 2020. Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations. *arXiv preprint arXiv:2002.01080*.

Srinivas, A.; Laskin, M.; and Abbeel, P. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*.

Tunyasuvunakool, S.; Muldal, A.; Doron, Y.; Liu, S.; Bohez, S.; Merel, J.; Erez, T.; Lillicrap, T.; Heess, N.; and Tassa, Y. 2020. dm_control: Software and tasks for continuous control. *Software Impacts*, 6: 100022.

Vamplew, P.; Dazeley, R.; Foale, C.; Firmin, S.; and Mummery, J. 2018. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1): 27–40.

Verma, M.; Bhambri, S.; Gupta, S.; and Buduru, A. B. 2019a. Making Smart Homes Smarter: Optimizing Energy Consumption with Human in the Loop. *arXiv preprint arXiv:1912.03298*.

Verma, M.; and Buduru, A. B. 2020. Fine-grained language identification with multilingual CapsNet model. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 94–102. IEEE.

Verma, M.; Kharkwal, A.; and Kambhampati, S. 2022. Advice Conformance Verification by Reinforcement Learning agents for Human-in-the-Loop. *arXiv preprint arXiv:2210.03455*.

Verma, M.; and Metcalf, K. 2022. Symbol Guided Hindsight Priors for Reward Learning from Human Preferences. *arXiv preprint arXiv:2210.09151*.

Verma, M.; Shah, N.; Nayyar, R. K.; and Hanni, A. 2021. Perfect Observability is a Myth: Restraining Bolts in the Real World.

Verma, M.; Sinha, P.; Goyal, K.; Verma, A.; and Susan, S. 2019b. A novel framework for neural architecture search in the hill climbing domain. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 1–8. IEEE.

Wilson, A.; Fern, A.; and Tadepalli, P. 2012. A bayesian approach for policy learning from trajectory preference queries. *Advances in neural information processing systems*, 25.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268.

Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; and Levine, S. 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, 1094–1100. PMLR.

Zahedi, Z.; Sreedharan, S.; Verma, M.; and Kambhampati, S. 2022. Modeling the Interplay between Human Trust and Monitoring. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1119–1123. IEEE.

Zahedi, Z.; Verma, M.; Sreedharan, S.; and Kambhampati, S. 2021. Trust-aware planning: Modeling trust evolution in longitudinal human-robot interaction. *arXiv preprint arXiv:2105.01220*.