

---

# Explaining Reinforcement Learning Policies through Counterfactual Trajectories

---

Julius Frost<sup>1</sup> Olivia Watkins<sup>2</sup> Eric Weiner<sup>3</sup> Pieter Abbeel<sup>2</sup> Trevor Darrell<sup>2</sup> Bryan Plummer<sup>1</sup> Kate Saenko<sup>1</sup>

## Abstract

In order for humans to confidently decide where to employ RL agents for real-world tasks, a human developer must validate that the agent will perform well at test-time. Some policy interpretability methods facilitate this by capturing the policy’s decision making in a set of agent rollouts. However, even the most informative trajectories of training time behavior may give little insight into the agent’s behavior out of distribution. In contrast, our method conveys how the agent performs under distribution shifts by showing the agent’s behavior across a wider trajectory distribution. We generate these trajectories by guiding the agent to more diverse unseen states and showing the agent’s behavior there. In a user study, we demonstrate that our method enables users to score better than baseline methods on one of two agent validation tasks.

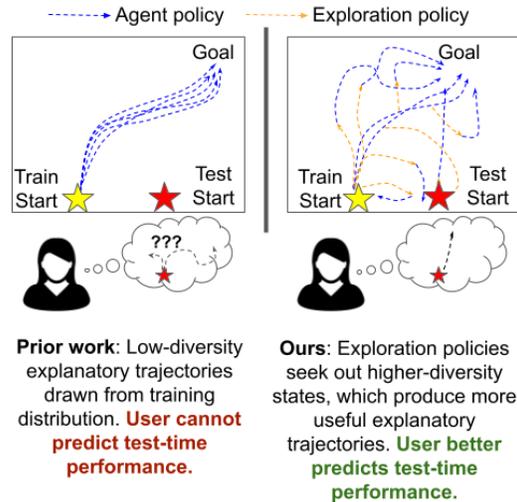


Figure 1. Our method seeks out more diverse states from which to show rollouts of agent behavior to the user.

## 1. Introduction

In most of the benchmarks where reinforcement learning (RL) has shown impressive results (Bellemare et al., 2013), (Tassa et al., 2018), agents are trained and tested in the same environment. In contrast, many practical applications involve sim-to-real transfer or a real-world train/test distribution shift which can hurt test-time performance (Eysenbach et al., 2020), (Akkaya et al., 2019), (Rusu et al., 2017), (Peng et al., 2018). Therefore it important to validate the performance and behavior of RL agents in the presence of a distribution shift. To do this, we take a step beyond summary metrics and focus on explainability techniques which help humans gain a more in-depth understanding of agent behavior.

Previous work in explainability often aims to help users

<sup>1</sup>Boston University <sup>2</sup>UC Berkeley <sup>3</sup>Harvey Mudd College. Correspondence to: Julius Frost <juliusf@bu.edu>, Olivia Watkins <oliviawatkins@berkeley.edu>.

understand and predict model behavior by explaining a set of model predictions (Hoffman et al., 2018). In the reinforcement learning setting, it may be desirable to select the most informative segments that show the behavior of the agent. With this motivation, critical states methods show the agent’s behavior in states which the agent thinks are important (Huang et al., 2018), (Amir & Amir, 2018). However, even if these methods successfully explain policy behavior, we cannot conclude if the behavior generalizes when the distribution shifts in the environment.

Our work builds upon the core idea of creating a set of informative trajectories of the agent’s behavior, but we focus on showing trajectories which are informative of the agent’s behavior under a test-time state distribution. To achieve this, we first define a prior over the types of states we expect to see at test time. Unfortunately, most simple priors (e.g. uniform) will also include many unreachable states which are not informative to show the user. To avoid this issue, we use an exploration policy to seek out states which match our prior distribution. By navigating to these new states rather than directly initializing our environment in these states, we reach a diverse yet still reachable set of states.

To summarize our contributions, we designed counterfactual trajectories that explain behavior policies in out-of-distribution states and conducted studies with human users that show increased understanding and generalization to out-of-distribution states by measuring prediction ability.<sup>1</sup>

## 2. Related Work

### 2.1. Saliency

One popular class of agent interpretability approaches is saliency methods, which aim to show which features from the input cause specific agent output. This method was employed by (Greydanus et al., 2017), (Hilton et al., 2020), and (Puri et al., 2020) to interpret which parts of the input that were deemed important to the agent’s decision. (Anderson et al., 2019) use saliency as a part of a more general agent explanation method. Even with all of the success, saliency methods have been shown to have limitations. (Atrey et al., 2020) show that saliency does not necessarily correspond to underlying representation in RL. In addition, saliency methods are not designed to deal with agents with memory, and in general may not be informative in a multi-timestep environment. All of these challenges point to saliency as a helpful explanatory tool but not as a stand-alone solution.

### 2.2. Interpretable Representations

Another approach is to use interpretable intermediate representations. In this approach, models are constructed so that even if they are not interpretable end-to-end, there will be some interpretable bottleneck or set of features used to make the final decision. (Madumal et al., 2019) use a causal model to generate explanations. (Chen et al., 2015) apply this method to autonomous driving, where a model is trained to take an image and output understandable intermediate representations in addition to a final action. Other work such as (Kim et al., 2018) and (Jiang et al., 2019) use language as an informative intermediate representation or output. While successful in some applications, these kinds of interpretability methods often rely on domain knowledge to choose appropriate intermediates, and they are not applicable if a policy’s decision cannot be explained by a small set of interpretable components.

### 2.3. Critical States

In the critical states framework, explanations are generated by showing especially informative “critical” states or trajectories from an agent. For example the work of (Huang et al., 2018) considers states critical if they have a large Q-value-difference between actions or a very low entropy

<sup>1</sup>Our code is available at <https://github.com/juliusfrost/cfrl-rllib>

action distribution under a maximum entropy learning regimen. (Amir & Amir, 2018) employ a similar approach, but they use videos instead of states and filter for diversity of trajectories. These methods can be seen as extensions over a baseline method of visualizing random states or failure states of an agent in an environment. The explanation format of our method - videos of the agent’s behavior in informative situations - is closest to this line of work.

### 2.4. Counterfactual Style Explanations

In (Rupprecht et al., 2019), researchers build a generative model of states and try to show users the agent’s behavior in states that optimize some notion of “interesting.” Again using a generative model, (Olson et al., 2019) use a GAN to produce modifications to states such that in the new state an agent takes a different action than it would have in the original state. While these works synthesize states directly, our method uses valid actions to access different parts of the state space. This means that any state we visit is actually reachable, at least in the training environment. The closest method to ours appears in the work of (Witty et al., 2018), in which the authors try to characterize generalization by exploring different starting configurations, either by directly changing the start state or by using states visited by another agent. In contrast to our work, their insights are applied to characterizing generalization, and not as explanations for humans to understand agent behavior.

(Anderson et al., 2019) provide a detailed empirical user study on the effectiveness of saliency map and reward explanations. They conclude there is no one explanation that fits all instances, but using several methods yields the best mental model, and thus encourage using multiple methods for explainability. We hope to introduce a novel explainability method which may add to this suite of tools.

## 3. Method

### 3.1. Preliminaries

In this work, we consider the challenge of generating informative trajectories which help a human user understand an agent’s policy (called the **behavioral policy**)  $\pi_\theta(a_t|s_t)$ , which is parametrized by weights  $\theta$  which can be trained through any RL or imitation learning method. The agent takes actions  $a_t$  in states  $s_t$ , resulting in a trajectory  $\tau$ .

Our goal is to help the user predict the agent’s behavior in a new test environment in the presence of a state distribution shift which results in different trajectory probabilities at train and test time:  $p^{\text{train}}(\tau|\theta) \neq p^{\text{test}}(\tau|\theta)$ . We consider two types of environment changes which can produce this distribution shift: differences in the initial state distribution ( $p_0^{\text{train}}(s_0) \neq p_0^{\text{test}}(s_0)$  for some initial state  $s_0$ ) and slight differences in the dynamics ( $p^{\text{train}}(s_t|s_{t-1}, a_{t-1}) \neq$

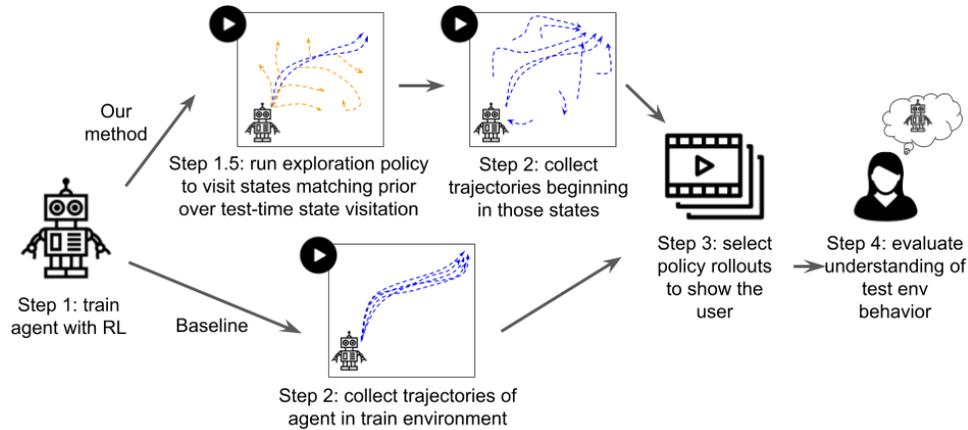


Figure 2. Explanation pipeline

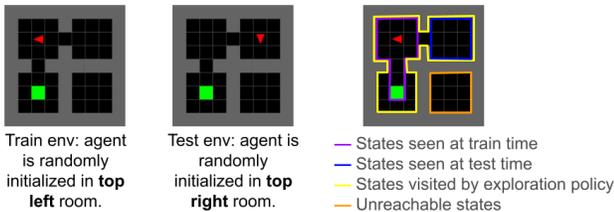


Figure 3. Distribution shifts, illustrated in the MiniGrid environment. Our user study was performed in an environment similar to this, except with the bottom-right room connected to the others.

$p^{\text{test}}(s_t | s_{t-1}, a_{t-1})$  for some state  $s_t$  and action  $a_t$ ). We assume that there are no changes to the agent’s action space or policy at test time - i.e.  $\pi_\theta(a_t | s_t)$  remains consistent between train and test time. We also assume that dynamics distribution shifts are minor. See Section 3.4 for more detailed discussion on this.

### 3.2. Method Overview

Figure 2 summarizes our method. We build upon the explanatory pipeline used in past critical states work such as (Huang et al., 2018). An agent is first pretrained using any RL or imitation learning algorithm. We then collect many trajectory rollouts of the agent, select a set to show the user, and finally test whether these explanatory trajectories have helped the human better predict the agent’s behavior in the test environment. This is challenging because of the distribution shift between the train and test environments.

Past work has attempted to select an informative set of trajectories to show the user by innovating on Step 3 in Figure 2 - the trajectory selection process. The critical states line of work selects trajectory segments which show the agent’s behavior at states where the policy believes one particular action is much better than the others (Huang et al., 2018).

Unfortunately, methods like these which sub-select trajectories from a set of agent rollouts in the train environment are often not very informative about the agent’s behavior under a distribution shift. Consider the train/test initial state distribution shift in Figure 3. A well-trained agent which was always initialized in the top-left room during training may take a nearly-deterministic trajectory navigating toward the goal every time, so none of the trajectories visited by the agent in the training environment will ever visit the states in the top-right room which the agent will experience at test time.

One way to generate trajectories in the training distribution which are more informative of the test distribution would be to manually change the starting state distribution. For instance, in the grid world shown in Figure 3, the simulator could be modified to directly initialize the agent in a uniform distribution over starting cells, ensuring that some trajectories will have been seen in whatever state the agent finds itself at test time.

One challenge with this approach is that it is often difficult to restrict the set of randomized initialization to reachable states the agent could possibly visit at test time. Instead, we propose creating a new distribution of start states using an **exploratory policy**  $\pi_\phi$  which can navigate from the states seen on the training distribution to a new, more diverse distribution of start states  $p_0^{\text{expl}}(s_0 | \phi)$  which are guaranteed to be reachable. Implementing this only involves adding one additional Step 1.5 to the pipeline shown in Figure 2.

Our trajectory generation method is summarized in the following algorithm:

1. Run the behavioral policy  $\pi_\theta$  in the train environment. Pause at a randomly selected state along the trajectory.
2. Run exploratory policy  $\pi_\phi$  starting at this state for a fixed number of timesteps. The exploratory policy

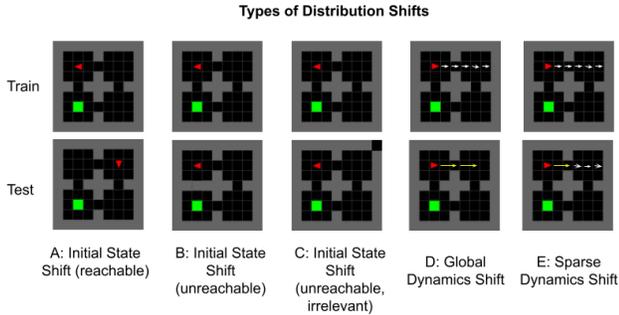


Figure 4. Possible types of train/test distribution shifts. Our user experiments use Shift A, although we would expect our method to be applicable to Shifts C and E as well.

will guide the agent to a set of states not typically observed in the training data. We call these **counterfactual states** because we can frame it as a counterfactual question: “what would the behavioral policy do if it instead navigated to this state?”

3. Run the behavioral policy starting from the counterfactual state for the remainder of the trajectory.
4. Show the user video rollouts of the agent’s behavior starting in these counterfactual states.

Note that the trajectories selection method used in item 4 of our algorithm can be very simple - in our experiments we randomly select trajectories. However, our method is complementary to other work on selecting informative trajectories, so future work could select trajectories more systematically, for instance filtering out near-repeated trajectories as was done in (Amir & Amir, 2018) or selecting based on critical states as was done in (Huang et al., 2018).

### 3.3. Exploration Objective

Our aim is to choose an exploration objective which will guide the agent to states which produce maximally informative trajectories. Intuitively, explanatory trajectories will be more informative the closer the distribution of trajectories visited by our exploratory policy is to the test-time distribution of trajectories visited. More formally, let  $p_0^{\text{expl}}(s_0|\phi)$  be the distribution of states in which the agent finds itself after running the exploration policy, and let  $p^{\text{expl}}(\tau|\phi, \theta)$  be the distribution of trajectories generated by rolling out the behavioral policy  $\pi_\theta$  initialized in  $p_0^{\text{expl}}(s_0|\phi)$ . Our goal is to achieve  $p^{\text{expl}}(\tau|\phi, \theta) \approx p^{\text{test}}(\tau|\theta)$ .

It is easiest to choose a principled exploration objective by making the assumption that train and test-time dynamics are approximately identical. In this situation, the train-test distribution shift consists entirely of changes to the initial state:  $p_0^{\text{train}}(s_0) \neq p_0^{\text{test}}(s_0)$ . However, if we are able to choose an

exploration policy which seeks out states matching the test time distribution - i.e. if  $p_0^{\text{expl}}(s_0|\phi) \approx p_0^{\text{test}}(s_0)$  - then because we roll out the same behavioral policy in environments with identical dynamics starting from nearly identical start distributions, we will achieve  $p^{\text{expl}}(\tau|\phi, \theta) \approx p^{\text{test}}(\tau|\theta)$ .

We do not precisely know  $p_0^{\text{test}}(s_0)$  before the agent experiences the test environment, so instead we choose a prior for our test time start state distribution. In the absence of domain knowledge about what states are likely under the test-time distribution shift, we assume a uniform prior over reachable test-time states and try to match this distribution as closely as possible by using an exploratory policy which seeks out a uniform distribution of states. (Note that with this exploration objective, the only reason we begin to run  $\pi_\phi$  along a trajectory generated by  $\pi_\theta$  rather than than running  $\pi_\phi$  from the start is to make achieving the exploration objective easier for a learned exploration policy.) However, if the user does have domain knowledge about which states are likely at test time, the exploratory policy could be modified to preferentially seek out those states. For instance, if the user believes that test-time states will be near the training distribution, the exploration objective may instead be to seek out a uniform distribution of states within a certain radius of those seen at train time.

There has been a variety of past exploration work in which agents are trained to seek out a uniform distribution of states or to learn distinct skills which take the agent to different parts of the state space such as (Sharma et al., 2019) and (Eysenbach et al., 2018). While any of these off-the-shelf exploration algorithms could be used, in our work, we experiment in the Minigrid environment, where the state space and dynamics are simple enough it is possible to hard-code an oracle exploration policy which achieves a uniform coverage over reachable states. By using this oracle policy, we are able to directly test the question of whether highly diverse states serve as better explanations rather than implicitly also testing the ability of the exploration policy.

### 3.4. Distribution shifts where Our method helps

There are multiple types of changes between train and test environments which can induce a difference between  $p^{\text{train}}(\tau|\theta)$  and  $p^{\text{test}}(\tau|\theta)$ , as illustrated in Figure 4. In Section 3.3, we gave an intuitive justification for why we would expect our method to help in cases where train and test-time dynamics are identical, and where the exploration policy is able to cover the test-time start distribution - Shift A, in Figure 4. Now, we consider our method’s usefulness in the presence of other distribution shifts.

Shift B starts the agent in a new state which is unreachable in the train environment, so it is unlikely that our explanations can help there. This restriction to only reachable distribution shifts is very limiting, but we can partially relax

it in cases where the distribution shift does not change the agent’s internal representation of the environment. Shift C illustrates an example of this: the test-time environment has the top-right corner wall segment removed, but so long as the agent’s internal representation of the environment does not change much based on this change, the agent will act the same way it would act if the wall was present. As a result, explanations generated in the train environment will still be informative about test-env behavior.

Finally, we consider dynamics shifts. In the case of global dynamics shifts, as shown in Shift D, our method is likely unhelpful, since train-time and test-time trajectories look very different even when starting from the same state. However, if dynamics shifts only occur in a few states in the test environment as shown in Shift E, then most of the time a train-time trajectory and a test-time trajectory beginning at the same state will look identical. Therefore, we can think of the dynamics shift purely as a way to introduce a state distribution shift. While our explanations will not help the user predict these dynamics shifts, the user should be able to predict how the agent will act after the dynamics shift leads the agent to an unseen state.

## 4. Experiments

In this section, we test our primary hypothesis that diverse states help the user understand the behavioral policy’s performance in the presence of a distribution shift. We empirically evaluate this through a Mechanical Turk user study for which we obtained IRB approval.

### 4.1. Study Design

Our study consisted of a questionnaire hosted online. Our participants were Mechanical Turkers from the United States with a high Mechanical Turk approval rating on past jobs. Each questionnaire consisted of an explanation phase in which the user sees video trajectories of the policy in the train environment and an evaluation phase which measures the user’s understanding of the policy’s performance in the test environment. Train and test environments differ in which room the agent is initialized, as shown in Figure 5. We tested participants on two tasks, described in Sections 4.1.2 and 4.1.3.

#### 4.1.1. EXPLANATION TYPES

For each of the tasks, our user study compares three different types of explanations. In the *Random States* setting, participants are shown 10 sample trajectories selected at random from a dataset of rollouts in an environment with the agent initialized in a particular start region. In the *Critical States* setting, participants are shown a trajectory containing each of the 10 lowest entropy states in the dataset, all with

the agent initialized in the same starting region each time. If two or more of the low-entropy states are in the same trajectory, the next lowest entropy states are selected. Finally, in the *Counterfactual States* setting, participants are shown 10 trajectories of the behavioral policy beginning from a state chosen by the oracle exploration policy, which seeks out a uniform distribution over reachable states. The *Random States* setting acts as a baseline for the other two methods, because it is common practice to look at random rollouts to understand policy behavior. The *Critical States* setting was chosen as another baseline, because it is the prior work with an explanation format which most closely matches our explanations, allowing for easier comparison. Finally, the *Counterfactual States* condition is our proposed explanation method. Since the agent is initialized in a different distribution of start states at test time, we expect the *Random States* and *Critical States* methods to be uninformative at helping the user predict test-time performance. In contrast, *Counterfactual States* show more diversity and should enable the user to predict test-time performance.

#### 4.1.2. BEHAVIOR UNDERSTANDING TASK

In this task, users first see videos of rollouts of  $\pi_\theta$  in the training environment selected using one of the three explanation methods described in section 4.1.1. In this task, as users watch these explanations, they are asked to build a mental model of the specific behaviors of  $\pi_\theta$ . They are then presented with a context state in the test environment, which is identical to the train environment except that the agent is initialized in a different part of the state space where  $\pi_\theta$  performs poorly. Below the context video, the user is presented with videos of three potential continuation trajectories from the context state. Only one of the continuation trajectories is generated by  $\pi_\theta$ , and participants must select the continuation corresponding to  $\pi_\theta$ . The incorrect choices for continuation trajectories were generated with policies manually designed to have behavior which is distinctly different from  $\pi_\theta$  - one of which is always successful in the test distribution, and one which fails in a way which is visually distinct from  $\pi_\theta$ .

#### 4.1.3. PERFORMANCE EVALUATION TASK

We next evaluate whether the user can predict the performance of  $\pi_\theta$  in the test environment. To do this, users are shown a context state from a test environment where the agent starts in a different state. The user must guess whether  $\pi_\theta$  will succeed or fail from that state, where success means the agent ends in the desired goal location. This task is designed to measure the participants’ understanding of the probability an agent succeeds in its task given its state.

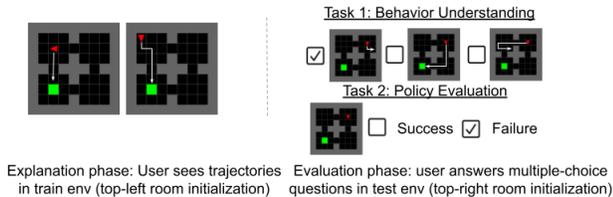


Figure 5. Experimental setup for the counterfactual states user study. Left: users see 10 train-time rollouts such as these in the explanation phase; Right: users are asked to either select the test-time behavior which corresponds with the policy they saw previously (Task 1 - Behavior Understanding) or predict the agent’s success in a new state (Task 2 - Policy Evaluation).

Table 1. Minigrid Policy Understanding and Evaluation

Condition	Task 1		Task 2	
	Accuracy	n	Accuracy	n
Random	0.04	10	0.6315	19
Critical	0.1458	10	0.6411	17
Counterfactual	<b>0.3638</b>	10	0.6904	21

Table 2. Task 1 is the Behavior Understanding task. Task 2 is Performance Evaluation. Bold means the result is statistically greater than the rest for  $p < 0.05$  with a one-sided T-test.

#### 4.1.4. EXPERIMENT SETTINGS

We test our method in a custom Minigrid environment (Chevalier-Boisvert et al., 2018) where there are four rooms with doors connecting between them (shown in Figure 4A). We train our behavioral policy using the A2C algorithm (Mnih et al., 2016) to good performance on the train environment. To select counterfactual states, we use an oracle policy which uniformly samples a feasible state and navigates towards it. In the study, each participant sees 10 explanations and then answers 10 questions of our chosen evaluation task.

For the Behavior Understanding task, the policy was trained in the top-left room, and explanations were generated in this room. In the test environment, the agent is positioned in the bottom-right room. As a result, the agent typically succeeds when initialized in the top-left (where we collect explanations) but typically fails when initialized in the top-right room (where we initialize in the evaluation phase).

For the Policy Evaluation task, we test users on two different behavioral policies. The first is only able to succeed starting in the top-left room, and the other is only able to succeed from the bottom-right room. In both cases, we collect explanations in the top-left room. Users are evaluated on the agent’s behavior in all four rooms.

## 4.2. Results

Qualitatively, in both tasks, the counterfactual states method produces more informative rollouts. In both tasks, neither the random states nor critical states methods produce any explanatory trajectories where the agent visits the room where the it is initialized at test time. This makes it hard for the user to predict the agent’s performance in these states. However, in the counterfactual states condition, the user sees multiple examples of the agent in the distribution of states it will see at test time.

Quantitative results, however, are mixed. Table 1 shows that for the Behavior Understanding task, users are able to select the correct behavior 36.38% of the time, a statistically significant improvement over the random and critical states conditions. This performance is only marginally above random guessing (33.3%), but at least the user did not seem to be actively misled by the explanations, as occurred in the random and critical states cases. On the Policy Evaluation task, users perform similarly in all conditions, with no statistically significant difference between them.

## 5. Discussion

While our method outperformed baselines in one of two tasks, there is significant room for improvement, as users did not consistently perform well in any condition of any task, even though the environment and agent behavior were both quite simple. Part of the difficulty in constructing good explanations derives from the population we were testing. Mechanical Turkers likely have little prior experience with reinforcement learning or time to carefully analyze agent behavior. Anecdotally, we also observed that many survey-takers come into the task with their own biases and preconceptions (for instance, assuming that a policy will be deterministic rather than stochastic, or assuming that if an agent does well in one situation it probably does well everywhere). Future work could test the usefulness of our counterfactual states method for agent developers who can spend time to familiarize themselves with the task, the explanation interface, and the types of behaviors which typically emerge in policies. In this setting, we could also test our method’s usefulness in more complex environments. The experiments we have run so far are quite limited: the environment is simple, an oracle exploration policy is available, the distribution shift only changes the agent’s starting position, and it is easy for a user to distinguish between correct and incorrect behaviors. Future work could explore whether counterfactual states hold value in more realistic settings.

In this paper, we hypothesized that showing higher-diversity reachable trajectories will enable users to better understand agent behavior. We tested this hypothesis through a user study in a gridworld in which the test-time initial state dis-

tribution differed from the train-time start distribution. Our method showed improvements over less diverse baseline methods in one of two tasks, but our experiments show room for improvement and illustrate the ongoing challenge of creating explanations which improve the understanding of non-expert users.

## 6. Acknowledgements

Thanks to Anna Rohrbach for ongoing feedback on this work.

## References

- Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik’s cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.
- Amir, D. and Amir, O. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1168–1176, 2018.
- Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Fern, A., and Burnett, M. Explaining reinforcement learning to mere mortals: An empirical study. *IJCAI*, 2019.
- Atrey, A., Clary, K., and Jensen, D. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep rl. *ICLR*, 2020.
- Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, jun 2013.
- Chen, C., Seff, A., Kornhauser, A., and Xiao, J. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pp. 2722–2730, 2015.
- Chevalier-Boisvert, M., Willems, L., and Pal, S. Minimalistic gridworld environment for openai gym. <https://github.com/maximecb/gym-minigrid>, 2018.
- Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- Eysenbach, B., Asawa, S., Chaudhari, S., Levine, S., and Salakhutdinov, R. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. 2020.
- Greydanus, S., Koul, A., Dodge, J., and Fern, A. Visualizing and understanding atari agents. *ICML*, 2017.
- Hilton, J., Cammarata, N., Carter, S., Goh, G., and Olah, C. Understanding rl vision. *Distill*, 2020. doi: 10.23915/distill.00029. <https://distill.pub/2020/understanding-rl-vision>.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.
- Huang, S. H., Bhatia, K., Abbeel, P., and Dragan, A. D. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3929–3936, Oct 2018. doi: 10.1109/IROS.2018.8593649.
- Jiang, Y., Gu, S., Murphy, K., and Finn, C. Language as an abstraction for hierarchical deep reinforcement learning. *arXiv preprint arXiv:1906.07343*, 2019.
- Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.
- Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. Explainable reinforcement learning through a causal lens. *AAAI 2020*, 2019.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.
- Olson, M. L., Neal, L., Li, F., and Wong, W.-K. Counterfactual states for atari agents via generative deep learning. *IJCAI*, 2019.
- Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.
- Puri, N., Verma, S., Gupta, P., Kayastha, D., Deshmukh, S., Krishnamurthy, B., and Singh, S. Explain your move: Understanding agent actions using specific and relevant feature attribution. In *ICLR*, 2020.
- Rupprecht, C., Ibrahim, C., and Pal, C. J. Finding and visualizing weaknesses of deep reinforcement learning agents. *arXiv preprint arXiv:1904.01318*, 2019.
- Rusu, A. A., Večerík, M., Rothörl, T., Heess, N., Pascanu, R., and Hadsell, R. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot Learning*, pp. 262–270. PMLR, 2017.
- Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.
- Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. DeepMind control suite. Technical report, DeepMind, January 2018. URL <https://arxiv.org/abs/1801.00690>.
- Witty, S., Lee, J. K., Tosch, E., Atrey, A., Littman, M., and Jensen, D. Measuring and characterizing generalization in deep reinforcement learning. *arXiv preprint arXiv:1812.02868*, 2018.