Towards Anomaly Detection in Reinforcement Learning

Blue Sky Ideas Track

Robert Müller
LMU Munich
Munich, Germany
robert.mueller@ifi.lmu.de

Steffen Illium
LMU Munich
Munich, Germany

Thomy Phan
LMU Munich
Munich, Germany

Tom Haider
Fraunhofer IKS
Munich, Germany

Claudia Linnhoff-Popien
LMU Munich
Munich, Germany

ABSTRACT

Identifying datapoints that substantially differ from normality is the task of anomaly detection (AD). While AD has gained widespread attention in rich data domains such as images, videos, audio and text, it has has been studied less frequently in the context of reinforcement learning (RL). This is due to the additional layer of complexity that RL introduces through sequential decision making. Developing suitable anomaly detectors for RL is of particular importance in safety-critical scenarios where acting on anomalous data could result in hazardous situations. In this work, we address the question of what AD means in the context of RL. We found that current research trains and evaluates on overly simplistic and unrealistic scenarios which reduce to classic pattern recognition tasks. We link AD in RL to various fields in RL such as lifelong RL and generalization. We discuss their similarities, differences, and how the fields can benefit from each other. Moreover, we identify non-stationarity to be one of the key drivers for future research on AD in RL and make a first step towards a more formal treatment of the problem by framing it in terms of the recently introduced block contextual Markov decision process. Finally, we define a list of practical desiderata for future problems.

KEYWORDS

Anomaly Detection; Reinforcement Learning; AI Safety

ACM Reference Format:
Robert Müller, Steffen Illium, Thomy Phan, Tom Haider, and Claudia Linnhoff-Popien. 2022. Towards Anomaly Detection in Reinforcement Learning: Blue Sky Ideas Track. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 5 pages.

1 INTRODUCTION

In Anomaly Detection (AD), one aims to identify datapoints that substantially differ from normality. AD is used to detect medical problems, fraud, production errors and many more. Since datapoints lack annotations of their degree of abnormality, the predominant approach is to learn a neural network (NN) based scoring function that models a notion of normality. Typical data sources in AD include images [34, 36, 43], video [30, 42], audio [29] and text [36]. Unlike in reinforcement learning (RL), these data sources are static and do not involve sequential decision making. In RL, an agent tries to maximize reward by interacting with its environment via trial and error. It is safe to assume that in realistic scenarios, trained RL agents will not exclusively act in the environment in which they were trained. Hence, unforeseen situations may arise such as the reduction of the braking power of an autonomous vehicle, people walking on the road, the appearance of a ghost driver, changing driving behavior of other traffic participants or the introduction of novel road signs. Since we cannot recreate all possible situations during training, the mismatch between training and deployment will be not the exception but the rule for real-world RL systems. Additionally, by acting in unfamiliar environments, the agent can also create accumulating anomalous situations through its own actions causing drastic changes it was not trained to account for. For example, an agent drives over a shelf because the shelf's position has changed, products fall off the shelf, block the way and damage the agent. Preventing such scenarios can be addressed by AD via early detection of possibly hazardous situations and is thus indispensable in safety critical scenarios. Refraining from AD based monitoring consequently bears far too many risks. However, we noticed a lack of research in the field of AD in RL. It is dominated by simple and unrealistic evaluation scenarios and does not fully embrace the unique challenges that arise due to the RL setting. This is partially due to the fact that it is less clear how to define and formalize the objective of AD in RL than for the other domains mentioned above. In this work, we aim to clarify the problem.

2 STATE OF RESEARCH

In this section we discuss related research and its limitations. Moreover, we outline similar questions and problems in other fields.

2.1 Simple Problems in Disguise

For a sound discussion of the challenges, limitations and open question that AD in RL entails, it is essential to review the current state of the field.

While concurrent work [18] has conceptually highlighted the importance of AD as one of the building blocks needed to enable safe RL systems, there is very little work that explicitly addresses anomaly detection in RL in terms of novel algorithms, domains or evaluation-scenarios. Moreover, we found that most existing works are rather traditional AD problems *in disguise*. Despite of claiming to tackle RL-specific AD problems, the proposed methods simply

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. © 2022 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

reduce to AD problems from classic Machine-Learning (ML) tasks, where RL merely serves as the data source. Additionally, approaches are usually evaluated on unrealistically simple scenarios.

For example, in [49] the authors aim to detect contaminated observations in a trajectory collected by an RL agent. Contaminated observations were constructed using additive Gaussian noise, simple white-box adversarial perturbations [13] or by replacing a random fraction of the observations in a trajectory with observations from another environment. The anomaly detector is trained on the observation's feature vectors (activations from penultimate layer of the agent's NN) based on the robust Mahalanobis distance and operates on a per-observation basis. This is basically the application of a traditional anomaly detector on a dataset of feature vectors. Moreover, the assumption that an adversary can directly modify the observations and propagate gradients through the agent to obtain high impact adversarial noise is unrealistic.

In [38], the authors use the entropy (measures uncertainty) of the predicted actions as an anomaly score to detect unencountered states. This is very akin to the simplest approach to out-of-distribution detection in classification [19] problems where one uses the entropy over the predicted classes. The method is trained on a fixed set of procedurally generated environments and tested whether it can detect other generated environments. In [37] the authors flip the observations vertically at test time.

Note that the methods do not account for the sequential nature of the observations. More importantly, the evaluation scenarios are artificial, simple and do *not* reflect realistic scenarios. Noise is very easy to simulate and, if desired, reliable methods to train noise robust NNs [10] exist. Detecting observations from deviating data-sources in the cases above results in visually very different observations (e.g. objects, colors, texture) which could very likely be detected using (pretrained) computer vision models. These problems are very similar to video or image AD.

The point is that these (simple) problems can easily be reduced to classic pattern recognition rather than representing unique AD tasks for RL and that the currently studied types of anomalies do not represent naturally occurring data.

Consequently, we need to look for more sophisticated AD challenges that embrace the RL setting and align better with real-world scenarios.

2.2 Related Problems

Having stressed the importance of finding suitable scenarios for AD in RL we extend our discussion with work that proposes interesting scenarios and approaches (mostly) without directly dealing with AD but could potentially benefit the field.

One line of work studies the out-of-distribution generalization when environment parameters are perturbed. Here the agent should be able to perform the task (or multiple tasks) in novel but related environments it has not encountered during training. This is important since it is well known that RL algorithms suffer from overfitting to the training environment [4, 48]. The considered scenarios include: a dexterous, simulated robotic hand that has to manipulate significantly smaller cubes [27] or additional objects [9] that it had not observed previously, the variation of an agent's limb length and width at test-time [28] or the modification of observations to have alternative background images or videos [17, 44].

Maximum entropy RL [9] and Self-Supervised Representation Learning [17] were shown to be particularly well suited to improve generalization. The former encourages exploration and helps to prevent premature convergence to sub-optimal policies while the latter optimizes an auxiliary objective that does not depend on external labels to construct a smooth and robust latent space that aids generalization. While these works are often only evaluated by changing the visual appearance of the observations, i.e., changing the MDP's state space, which then essentially amounts to facilitating domain adaption approaches from computer vision, some of these works also change the agents morphology and mechanics during evaluation [24, 28] or directly aim to close the simulation-to-reality gap. This results in a change to the MDP's dynamics as well as the state-space [16].

Whether methods that improve generalization can still detect if the agent is facing novel and previously unencountered situations or if these methods deteriorate the detection performance is an intriguing question that remains to be answered.

Although the evaluation scenarios above are used to analyze the generalization capabilities of different algorithms, we argue that this is also a challenging and more realistic setting to evaluate AD in RL since it is clearly more specific to the RL setting. To the best of our knowledge, the authors of [7] are the first to make a step in this direction and coin the problem Out-of-Distribution Dynamics Detection (OODD). Hence, the adjusted environment dynamics are interpreted as anomalies that need to be detected within a short time frame. They propose four different types of anomalies that can occur: iid noise (apply gaussian noise), sensor shutdown (set some features to zero), sensor calibration failure (multiply features by a constant) and sensor drift (increase magnitude of noise over time). Note how these anomalies are solely based on noise in environments with continuous, low-dimensional observations. The first three types of anomalies are often directly robustified against using robust RL approaches [9, 33] or simple data augmentations [22] as they are easy to anticipate and lie in a region where detection might not even be worthwhile.

On the other hand, sensor drift is a *non-stationary* anomaly source. We believe that non-stationarity is one of the key challenges future AD system should be able to deal with in the context of RL as the world is (from our point of view) inherently non-stationary.

Related to that is the problem of lifelong (continual) reinforcement learning [20]. Here the agents needs to learn and adapt in a constantly changing environment and tasks. Hence, there is no clear separation between training and testing. However, the problem itself is motivated by non-stationarity where one distinguishes between active and passive non-stationarity. When the agent can influence the nature of the non-stationarity through its own behavior (e.g. hallucinating new goals) one speaks of active and else of passive non-stationarity.

We argue that in order to make progress and to create new challenges, we need to come up with settings where anomalies are more subtle, semantically more meaningful and not simple, easy-to-specify augmentations of the state space and/or the transition function. Anomalies must be sufficiently complex such that they cannot be anticipated beforehand. Ultimately, the anomaly source itself could be governed by another RL agent. Inevitably, this will result in less general evaluation scenarios as complex anomalies are domain-dependent.

3 THE GENERALIZATION-DETECTION DILEMMA

The previous section made it obvious that while generalization in RL follows a very different agenda than AD, both share a similar setup at their core.

A typical approach to enhance generalization is to make the agent invariant to task-irrelevant properties of the environment as they are assumed to represent a source of distraction for the RL algorithm, e.g. an autonomous car on the highway that does not care about the changing surroundings (e.g. landscape, city, weather) through which it travels as it approaches its destination.

In domain randomization one simply trains on enough variations of the environment [6]. This approach suffers from high sample complexity and relies upon a simulator that can produce a diverse set of environments. Similar approaches use data augmentations [23, 44], to artificially create "new" environments. Others [11, 47] use bisimulation [12] metrics where two states are bisimilar if they have similar long-term behavior, i.e. similar expected rewards and dynamics.

We believe that the notion of ignoring everything that is not related to the task is fine in controllable simulations but might be detrimental in more safety-critical scenarios such as autonomous driving or smart factories [32]. The same holds for blindly adapting to novel situations and effectively making them in-distribution, for example by optimizing auxiliary-objectives (e.g. rotation-prediction) at test-time using gradient descent [17]. In safety-critical scenarios, one typically deals with multiple tasks and safety constraints. *Designing a suitable objective in such complicated scenarios is usually hard if not infeasible. Specification flaws can lead to pathological situations where the agent retains its capabilities out-of-distribution yet pursues the wrong objective [21]. Therefore, ignoring information that is irrelevant to a misaligned objective (even if only for a small amount) poses a safety critical threat. Hence, we identify AD components in RL systems to be of particular importance in safety critical domains and advocate for more engagement with the topic in the RL community.*

Imagine an autonomous car whose only goal is to drive from *a* to *b* on the highway. It might ignore signs of emerging forest fires in the surrounding landscape. An AD system on the other hand, could detect the threat and propagate it to the operator for further instructions. *Finding a suitable response strategy to react to detected anomalies yields another crucial aspect to consider for future AD systems in RL,* e.g. by switching to a more conservative or specialized policy or handing over control to a human. In a lifelong learning scenario, the agent could decide whether to keep adapting or not based on an anomaly score. Additionally, *one should minimize false-positives to combat alert-fatigue [3] and frequent policy switching. However, note how there is a thin line between making the agent robust or invariant to small changes in the environment and the need to detect anomalies. What is meant by "small changes"? What should be detected, what can safely be ignored?* We conjuncture that it is best to combine the best of both worlds. The agent should have sufficient generalization capabilities such that it can act sensibly under the influence of novel dynamics and observations while still being able to detect and report these situations as anomalous. Domain knowledge can be used to specify irrelevant factors of variation ("small changes"). The latter is addressed by practitioners through the use of operational design domain specifications to explicitly define to what extent the agent must be safe to operate, i.e., autonomous driving regardless of the weather or traffic. In the future, less invasive approaches are desirable.

A common misconception is to assume that all anomalies are dangerous. This decision depends on the agent's notion of safety. However, without the ability to detect anomalies, this decision cannot be made.

4 A MORE FORMAL VIEW

The previous discussion lacks a formal description of the task of AD in RL. In this section, we aim to connect the problem with a variant of an MDP that incorporates the aspect of slowly drifting dynamics, the block contextual Markov decision process (BC-MDP) [40]. Compared to other formalism such as the hidden parameter block MDP [47] or the dynamic parameter MDP [45] it does not rely on an episodic setting. In particular, dynamics are allowed change within episodes and not only across episodes. In safety-critical scenarios we must be able to detect changes in a matter of timesteps and not episodes (if any).

The following definition of a BC-MDP is taken from [40].

A BC-MDP is defined by a tuple $(\mathcal{C}, \mathcal{S}, \mathcal{A}, \mathcal{M})$ where $\mathcal{C}$ is the context space, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space and $\mathcal{M}$ is a function which maps a context $c \in \mathcal{C}$ to MDP parameters and observation space $\mathcal{M}(c) = \{T^c, S^c\}$.

In short, state space and transition dynamics are governed by context c.

For the AD setting, let $p(c)$ be the density of some $c \in \mathcal{C}$ and let $C_{normal}, C_{anomalous} \subset \mathcal{C}$. Further, we assume that

$$C_{normal} \cap C_{anomalous} = \emptyset \ \wedge \ C_{anomalous} = \mathcal{C} \setminus C_{normal}$$

Then the goal is to find a suitable p such that

$$p(c) > p(c') \quad \forall \, (c, c') \in C_{normal} \times C_{anomalous}$$

In practice we are usually restricted to some $C'_{normal} \subseteq C_{normal}$ where the context is not directly observable and needs to be inferred (under the assumption of identifiability). That is, we do not know the context space and are given $\mathcal{M}(c)$ but not c. Hence, we frame the problem of AD in RL as detecting whether the agent is acting in a familiar or unfamiliar context[1].

5 PRACTICAL DESIDERATA FOR FUTURE PROBLEMS

We now define various desiderata for future AD problems we believe to be both realistic and practical.

I. Sufficient data to model normality is available

Access to enough data to model normality is crucial and a common assumption in AD problems. One assumes that normal data is cheap and easy to acquire whereas retrieving anomalous data is scarce, expensive to obtain, impractical and in most cases simply impossible. Therefore we cannot expect to get an exhaustive set of anomalous data. This is the very reason we cannot rely upon supervised learning (SL). In the context of RL this could mean having access to a few training tracks (or levels) where one can control for

[1] For simplicity we do not include the case where C'_{normal} is contaminated by a small fraction of anomalous contexts but the above could easily be extended to do so.

normality, e.g. not having humans cross the street during training.

II. Anomalies are semantically rooted in the environment
We have already discussed this point in depth in the previous sections. We need to move away from simple pattern recognition problems towards studying semantically meaningful anomaly sources that are deeply rooted in the environment under inspection and exhibit non-stationarity. E.g. the occasional presence of road-workers or unseen traffic signs at test time. A similar trend can be observed in computer vision [8]. Moreover, one might also introduce anomalies into the non-stationarity itself, i.e., by varying the degree of non-stationarity.

III. Absence of rewards during deployment
This point is akin to the problem of scalable supervision [1]. In real-world scenarios it is often infeasible to evaluate (compute reward) the agent's actions, for example when it requires human supervision or ties up valuable resources such as external monitoring systems. Therefore the absence of reward during deployment is more realistic. When one is only interested in the deterioration in performance with respect to the reward and the reward is available at test-time, statistically comparing rewards from train and test-time [14] yields good results. We exclude this setting.

IV. Simultaneous or post-hoc,
whitebox or blackbox AD-training
Regarding the training of dedicated AD modules we are faced with several choices that arise due to the RL setting: (i) Train the detector in conjunction with the agent. The challenge is to account for the agent's learning dynamics. Note that it is possible to use the inherent properties such as the uncertainty over actions and a dedicated AD might not be needed. (ii) Training the detector post-hoc. Here we need generate additional rollouts with the trained agents in the train environment. The challenge is to determine how many rollouts are needed to get enough data to model normality and how to do this efficiently. The post-hoc method introduces a source of redundancy and increases sample complexity which can be problematic when rollouts are expensive. However, it might alleviate the problem of having to deal with the learning dynamics of the agent. (iii) Blackbox training. In this setting one can only observe the data emitted by the agent, i.e. states, actions, and next states. In particular we do not have access to any internal states such as the activations of the agent's NN. For example, a robot manufacturer may choose not to allow access to the NN as this would give away their unique selling point. (iv) Whitebox training. As opposed to (iii), here we have access to all parts (e.g. its activations) of the agent and are also allowed to add additional training phases and objectives, e.g. to add auxiliary losses to improve detection performance.

V. Minimize false positives/negatives and detection lag
We should aim for a low false positive rate to ensure regulated and practicable operation and combat alert fatigue. Minimizing false negatives ensures that we do not overlook anomalies. Finding the right balance is a key challenge. Moreover, it is important to quickly detect anomalous situations from as few samples or interactions as possible, i.e. we should minimize the detection lag.

VI. Design a suitable response strategy
Assuming we have a good anomaly detector, the next question that arises is how we handle situations that are flagged as anomalous (ideas discussed in Section 3). This question is presumably very domain depended and bears its own challenges and can also be studied in isolation. Nevertheless, when deploying agents in the real world this question is indispensable.

6 FUTURE APPROACHES

We will now briefly outline some possible solutions that we believe should be further explored in the future.

One obvious paths forward is to study how and to what extend exploration approaches can be re-purposed at deployment to discover anomalous states (or situations). One could for example use the entropy of random state-features [39], the distillation error when regressing the features of another random network [5], the uncertainty over intermediate actions from an inverse curiosity module [31] or density based pseudo counts [50]. However these methods only explore the consequences of short-term decisions.

Model based RL (MBRL) tries to explicitly construct a model of the components of the MDP. One could use these models to check whether the received observations and transitions align with the learned model of the world [15], e.g. by comparing the predictions with the actual outcome. Additionally, with the access to anomalous situations one could try to hallucinate [2] novel anomalous situations using the agent's world model. Yet, MBRL approaches are still very brittle and hard to optimize.

Another promising direction is to directly infer *normality contexts*. In computer vision, the idea of learning representations using self-supervised learning [41, 43] without external supervision and applying AD on these representations has recently gained traction. One could equip the agent with similar representation learning capabilities [26] to infer a sensible space of context representations during training. From there on, the space of normal contexts can be made more compact [34, 35] in a post-hoc fashion such that anomalous contexts will be placed farther from the center in representation space. Repurposing approaches that deal with concept-drift and concept-shift in time-series could be another fruitful avenue. Here the challenge is to incorporate the decision making component.

Lastly, by inferring the environment's causal structure [25, 46] we might gain more insights into what caused the anomaly and use it to make AD more explainable.

7 CONCLUSION

In this paper we discussed the meaning of anomaly detection in the context of RL and identified that current approaches and evaluation scenarios are insufficiently realistic. We linked the problem to various other topics in RL such as lifelong learning and generalization and gave concrete instructions on how to design more realistic scenarios in the future. Additionally we made first steps towards the formalization of the problem and discussed various ideas on how future work could approach the problem. AD in multi agent RL is another frontier which adds another dimension of complexity. Anomalies may arise in the communication between agents, in their ability to cooperate or through the introduction of adversaries.

ACKNOWLEDGMENTS

This work was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv:1606.06565* (2016).

[2] Philip Ball, Cong Lu, Jack Parker-Holder, and S Roberts. 2021. Augmented World Models Facilitate Zero-Shot Dynamics Generalization From a Single Offline Environment. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*.

[3] Tao Ban, Ndichu Samuel, Takeshi Takahashi, and Daisuke Inoue. 2021. Combat Security Alert Fatigue with AI-Assisted Techniques. In *Cyber Security Experimentation and Test Workshop*. 9–16.

[4] Emmanuel Bengio, Joelle Pineau, and Doina Precup. 2020. Interference and generalization in temporal difference learning. In *International Conference on Machine Learning*. PMLR, 767–777.

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random network distillation. In *International Conference on Learning Representations*.

[6] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*. PMLR, 2048–2056.

[7] Mohamad H Danesh and Alan Fern. 2021. Out-of-Distribution Dynamics Detection: RL-Relevant Benchmarks and Results, In International Conference on Machine Learning, Uncertainty & Robustness in Deep Learning Workshop.

[8] Lucas Deecke, Lukas Ruff, Robert A Vandermeulen, and Hakan Bilen. 2021. Transfer-based semantic anomaly detection. In *International Conference on Machine Learning*. PMLR, 2546–2558.

[9] Benjamin Eysenbach and Sergey Levine. 2021. Maximum entropy rl (provably) solves some robust rl problems. *arXiv:2103.06257* (2021).

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

[11] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. 2019. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*. PMLR, 2170–2179.

[12] Robert Givan, Thomas Dean, and Matthew Greig. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147, 1-2 (2003), 163–223.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv:1412.6572* (2014).

[14] Ido Greenberg and Shie Mannor. 2021. Detecting Rewards Deterioration in Episodic Reinforcement Learning. In *International Conference on Machine Learning*. PMLR, 3842–3853.

[15] David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2455–2467.

[16] Tom Haider, Felippe Schmoeller Roza, Dirk Eilers, Karsten Roscher, and Stephan Günnemann. 2021. Domain Shifts in Reinforcement Learning: Identifying Disturbances in Environments. (2021).

[17] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. 2020. Self-Supervised Policy Adaptation during Deployment. In *International Conference on Learning Representations*.

[18] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved Problems in ML Safety. *arXiv:2109.13916* (2021).

[19] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

[20] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. 2020. Towards continual reinforcement learning: A review and perspectives. *arXiv:2012.13490* (2020).

[21] Jack Koch, Lauro Langosco, Jacob Pfau, James Le, and Lee Sharkey. 2021. Objective Robustness in Deep Reinforcement Learning. *arXiv:2105.14111* (2021).

[22] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement Learning with Augmented Data. *Advances in Neural Information Processing Systems* 33 (2020).

[23] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. 2019. Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning. In *International Conference on Learning Representations*.

[24] Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, and Jinwoo Shin. 2020. Context-aware dynamics model for generalization in model-based reinforcement learning. In *International Conference on Machine Learning*. PMLR, 5757–5766.

[25] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems* 33 (2020), 11525–11538.

[26] Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. 2021. On The Effect of Auxiliary Tasks on Representation Dynamics. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1–9.

[27] Daniel J Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Yuanyuan Shi, Jackie Kay, Todd Hester, Timothy Mann, and Martin Riedmiller. 2019. Robust Reinforcement Learning for Continuous Control with Model Misspecification. In *International Conference on Learning Representations*.

[28] Khushdeep Singh Mann, Steffen Schneider, Alberto Chiappa, Jin Hwa Lee, Matthias Bethge, Alexander Mathis, and Mackenzie W Mathis. 2021. Out-of-distribution generalization of internal models is correlated with reward. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*.

[29] Robert Müller., Steffen Illium., Fabian Ritz., and Kyrill Schmid. 2021. Analysis of Feature Representations for Anomalous Sound Detection. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. INSTICC, SciTePress, 97–106.

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14372–14381.

[31] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.

[32] Thomy Phan, Lenz Belzner, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, and Claudia Linnhoff-Popien. 2021. Resilient Multi-Agent Reinforcement Learning with Adversarial Value Decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11308–11316.

[33] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2817–2826.

[34] Tal Reiss and Yedid Hoshen. 2021. Mean-Shifted Contrastive Loss for Anomaly Detection. *arXiv:2106.03844* (2021).

[35] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *International Conference on Machine Learning*. PMLR, 4393–4402.

[36] Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. 2019. Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. 4061–4071.

[37] Andreas Sedlmeier, Thomas Gabor, Thomy Phan, and Lenz Belzner. 2020. Uncertainty-Based Out-of-Distribution Detection in Deep Reinforcement Learning. *Digitale Welt* 4, 1 (2020), 74–78.

[38] Andreas Sedlmeier, Robert Müller, Steffen Illium, and Claudia Linnhoff-Popien. 2020. Policy Entropy for Out-of-Distribution Classification. In *International Conference on Artificial Neural Networks*. Springer, 420–431.

[39] Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2021. State Entropy Maximization with Random Encoders for Efficient Exploration. *arXiv:2102.09430* (2021).

[40] Shagun Sodhani, Franziska Meier, Joelle Pineau, and Amy Zhang. 2021. Block Contextual MDPs for Continual Learning. *arXiv:2110.06972* (2021).

[41] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. 2020. Learning and Evaluating Representations for Deep One-Class Classification. In *International Conference on Learning Representations*.

[42] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.

[43] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. 2020. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. *Advances in Neural Information Processing Systems* 33 (2020), 11839–11852.

[44] Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. 2020. Improving Generalization in Reinforcement Learning with Mixture Regularization. In *NeurIPS*.

[45] Annie Xie, James Harrison, and Chelsea Finn. 2021. Deep Reinforcement Learning amidst Continual Structured Non-Stationarity. In *International Conference on Machine Learning*. PMLR, 11393–11403.

[46] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. [n.d.]. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*.

[47] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2020. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations*.

[48] Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. 2018. A study on overfitting in deep reinforcement learning. *arXiv preprint:1804.06893* (2018).

[49] Hongming Zhang, Ke Sun, Bo Xu, Linglong Kong, and Martin Müller. 2021. A Simple Unified Framework for Anomaly Detection in Deep Reinforcement Learning. *arXiv:2109.09889* (2021).

[50] Rui Zhao and Volker Tresp. 2019. Curiosity-driven experience prioritization via density estimation. *arXiv:1902.08039* (2019).