Measures of Generalisation for Deep Reinforcement Learning

Patrick Leask, Mark Flowers, Arpita Goyal, Saeth Wannasuphoprasit,
Akshay Shrinivas Munde, Kshitij Singh, Amol Sahebrao Rathod

1 Introduction

Deep reinforcement learning has showed substantial promise at solving tasks with highly dimensional inputs, showing better-than-human performance at games such as Go [SSS+17] and StarCraft II [VBC+19]. As these models are employed in increasingly safety critical applications, the need to provide guarantees on agents' performance in unseen environments becomes more important [JN+19]. In order to provide these guarantees, and to build increasingly generalisable models, we need reliable and comparable measures for generalisation.

In this paper, we seek to provide a holistic overview to current approaches to studying the generalisation performance of deep reinforcement learning models. In 2 we justify the importance of studying generalisation in the specific context of deep reinforcement learning models. 3 sets the problem and notation for the discussion in 4 of different approaches to studying generalisation.

2 Generalisation and Intelligence

Generalisation is a quality of intelligent behaviour and can be thought of as environmental adaptivity. Intelligent systems navigate obstacles, cope with changing sensory input and negotiate environmental perturbations all of which threaten to stop them achieving their goals. Commonplace human capacities are extraordinarily flexible. Consider walking: human beings fluidly cope with changes in incline, different surfaces, stationary and moving obstacles, and we can do all of this while engaged in conversation or looking at a phone. There is a broad consensus that the kind of generalisation we are describing is a necessary feature of intelligent behaviour [LH+07].

Generalisation is inextricable from intelligence, so it is crucial that models aiming to replicate or exceed human intelligence can continue to succeed at their tasks in the face of environmental variations. Numerous studies have appeared that focus on ways to create models that exhibit generalisation (see [ZM+21] [KH+21] inter alia). So far we have discussed generalisation conceptually, but it needs to be framed as a quantity to be a useful tool in research. Henderson et al [HI+18] have pointed to the need for standardised measures of performance for DRL models. With different research teams using idiosyncratic performance metrics it is unclear if everyone is measuring the same quantity. There is a similar gap in current research when it comes to measuring generalisation. Our work aims to provide a holistic overview of generalisation measures, with a focus on applying them, rather than a theoretical analysis of measures which has been done elsewhere [JN+19].

Deep Reinforcement Learning (DRL) involves leveraging deep neural networks to find a reward-maximising policy for behaviour. The concepts of DRL are discussed in detail in 3. Researchers working on DRL have postulated that reward maximisation is sufficient for intelligence [SSPS21]. These high-hopes for DRL are based on models achieving super-human performance in different tasks by developing policies that maximise reward [MKS+15] [SSS+17] [VBC+19]. We have chosen to look at generalisation in DRL models in particular because there is strong research interest in the area, with world-leading AI research teams believing that DRL will produce general artificial intelligence [SSPS21].

Although DRL models have achieved super-human capacities for various tasks [MKS+15] [SSS+17] [VBC+19], they have yet to match human-level generalisation [GTR+18]. On an object recognition tasks humans outperform DRL models when faced with recognising an object in an image under different modes of degradation. Interestingly, DRL models can achieve super-human performance when trained on images that are distorted in a specific way but this training does not significantly improve their object recognition ability when faced with images subjected to a different form of distortion. Human subjects handle the change in type of distortion fluidly, demonstrating consistent performance in the object recognition task across 12 types of image distortion.

Generalisation is a topic of interest for any approach to AI, but it is of particular interest with regard to DRL models as they are vulnerable to over-fitting to the environment they are trained in [ZV+18] [WT+11]. Overfitting to a training environment can be understood as a generalisation failure: the model excessively adapts to the peculiarities of the environment it is trained in, preventing it from performing a task successfully in previously unseen environments.

3 A Brief Introduction to Deep Reinforcement Learning

In this section we give background for Deep Reinforcement Learning (DRL). Due to space constraints, and to stay focused on our main topic, measures of generalisation for DRL models, we omit some of the formal and technical details. References for further reading are provided.

3.1 Deep Learning

Deep learning models are a class of machine learning models that build abstract input data into multiple levels of higher-order representations. Similarly to how the human eye pre-processes raw input into basic shapes, which are pieced together by the brain; a deep learning model learns to build its own internal representations to better understand the data [LBH15]. In contrast, traditional machine learning techniques often require engineered feature vectors as input [LBD+89].

The ability of deep learning models to build their own representations of highly dimensional data makes them well suited for tasks such as image recognition [KS+12] and, as we will see, reinforcement learning.

3.2 Reinforcement Learning

Reinforcement learning is essentially learning how to behave over time in order to get the best reward. The learner is referred to as the agent who, over time, interacts with an environment, which is everything other than the agent. For each action the agent takes, the environment provides a numerical reward. This feedback cycle is represented in 1.

Precisely, we say that over a series of discrete time steps $t = 0, 1, 2, ...$, an agent receives information about the environment's state $s \in S$, where S is the set of all possible states. Based on this state, the agent chooses an action $a_t \in A(s_t)$, where $A(s_t)$ is the set of actions available in state s_t. At the next time step, the agent receives a numerical reward $r_{t+1} \in R$ where $R \subseteq \mathbb{R}$ and is in a new state s_{t+1}. This transition may be either deterministic or stochastic, in which case the likelihood of an action a in state s resulting in a transition to s' is given by a transition function $T : S \times A \times S$. Often the agent is unable to observe the entire state space, in which case the agent makes an observation $\omega \in \Omega$.

The agent's chosen action is determined by a policy π_t, which maps states to probabilities of selecting an action at each time step, where $\pi_t(s, a)$ is the probability of the agent selection action a at time step t in state s. For a dataset D obtained on this task, consisting of tuples $\langle s, a, r, s' \rangle \in S \times A \times R \times S$, a learning algorithm is a function $f : D \to \pi_D$.

The agent seeks to maximise its expected return $V^{\pi}(s) : S \to \mathbb{R}$, such that:

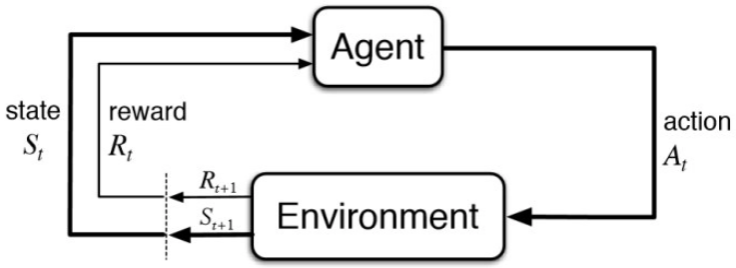

Figure 1: The agent-environment interaction in reinforcement learning [SB98].

$$V^{\pi}(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, \pi]$$

Where $r_t = \mathbb{E}_{a \sim \pi(s_t,.}R(s_t, a, s_{t+1})$, and γ is a discount function applied to the reward at each step.

A more in-depth formulation of the reinforcement learning problem can be found in [SB98].

3.3 Deep Reinforcement Learning

Deep reinforcement learning is the application of deep learning models to the reinforcement learning problem, ie. where the learning algorithm $f : D \to \pi_D$ leverages a deep neural network. This combination has seen particular success in tasks with highly dimensional input spaces, such as the board game Go [SSS+17].

4 Measures of Generalisation

Measures of generalisation are quantities that represent a model's ability to successfully perform a task in the face of changing environmental conditions - these quantities might be determined by performance on a benchmark test, or be determined using formal properties of the model. We focus on measures of generalisation where the task is consistent and the environment changes. One might also attempt to measure a model's ability to transfer learning on one task in a fixed environment to a similar task conducted in the same environment. Here we are concerned only with the former notion of generalisation. We look at empirical measures, which aim to design benchmark tests to 'score' a model's ability to generalise, and theoretical measures, which quantify generalisation using the intrinsic formal properties of a model.

4.1 Empirical Measures

In this section we discuss approaches that quantify a model's ability to generalise by evaluating its online performance against a generalisation benchmark. In particular, we discuss an approach that compares performance in a training environment to performance in procedurally generated environments [CK+19] [ZV+18] and an approach that measures whether performance co-varies with changes to environmental parameters [PG+18]. Procedurally generated environments are created using an algorithm that takes a random seed as input and outputs an environment, in the case of [CK+19] these are 2D spaces with platforms of various elevations and a range of obstacles.

The CoinRun environment [CK+19] is a game in which the agent controls a character that must navigate a character through a 2D space to a coin. The coin is on the far side of the space from the character and various obstacles. Colliding with an obstacle results in failure and reaching the

coin results in reward. An example is shown in 2. Cobbe et al found that as the number of levels models were trained on increased performance on unseen test levels approached performance on training levels - where performance is understood as the percentage of levels in which the character reaches the coin. An agent's ability to generalise is measured by the discrepancy between training and test performance.

The environmental parameters approach involves training an agent in an environment that has several parameters that can be varied and then testing the agent against increasingly extreme variations of those parameters. Packer et al [PG+18] use a collection of six such tests to benchmark an agent's ability to generalise. One of these is the MountainCar task: the agent must move a (simulated) car up a hill, pushing the car left or right at each time step; two environmental parameters can be varied, push force magnitude and car mass. Performance in the training environment and the test environment is measured in terms of the percentage of runs in which the car was moved to the top of the hill. The agent's ability to generalise is again understood as the discrepancy, or lack thereof, between training performance and test performance. There is an added level of nuance here over the CoinRun benchmark because the extent to which the test environment differs from the training environment can be measured in terms of how extreme the variation of environmental parameters is. In other words, a 'distance' from the training environment within which the agent can succeed at the task can be determined.

Both the procedural generation approach to empirical benchmarking of generalisation and the environmental parameters approach have the advantage of being model agnostic. Researchers have designed the benchmarks to be tractable for a range of DRL algorithms [CK+19] [PG+18]. One drawback of standardised benchmarks is that researchers might overfit the architecture of their models to such a benchmark. Furthermore, the environments used in the CoinRun benchmark, although procedurally generated, have limited variability. The tests used in [PG+18] likewise allow for variation of just a few environmental parameters. The real world has many more parameters which can vary in unexpected ways, so these tests arguably do not show that a model can generalise in real world environments. One of the motivations for using DRL is that it can handle higher-dimensional problems, but the environmental inputs in the tests we have reviewed are relatively sparse.

Part of the motivation for the simplicity of these benchmarks is to make them tractable for currently available DRL models and learning. We suggest that a suitable benchmark for generalisation should be scalable. As DRL models become more powerful a suitable benchmark should be able to increase the variance in environments the model's generalisation is tested against. Scalability would also help to address the potential overfitting of model design to benchmark tests. Since the test environment of a scalable benchmark would not be fixed a model's generlisation would be determined by the range of variations that can be made before it no longer performs its task. In other words, generalisation would be understood as a range of environmental changes within which the model's performance remains robust. Even if a model were overfitted by design to a certain range of scalable test environments, that would not necessarily mean that the model was overfitted to the entire range of possible test environments.

A less tractable issue for the empirical approach to measuring generalisation is explainability. Success on a generalisation benchmark does not tell us why a particular model was able to generalise to the range of test environments it was successful in. Ideally, a generalisation benchmark would make it clear why a model can generalise and thereby inform future model design.

4.2 Theoretical Measures

Empirical assessment of a model alone cannot complete our understanding of why a certain model generalises well, it can simply provide us experimental information about which architectures and hyperparameter values correlate well to generalisation performance. Furthermore, it may not always

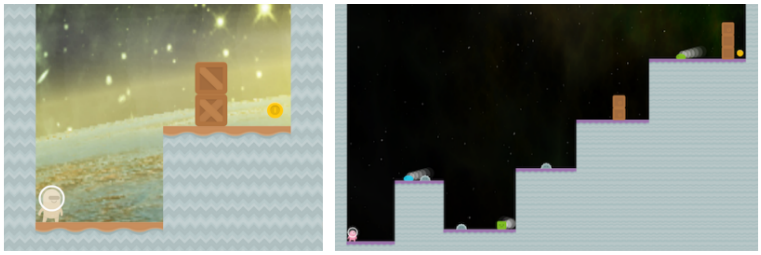

Figure 2: Two levels from CoinRun [CK+19].

be possible or appropriate to provide further examples outside of the training dataset on which to evaluate the model performance. Based on the model, the optimiser, and the nature of the data, we seek to determine what causes a model to generalise well. The structure of the samples from the generative distribution is of fundamental importance to this problem, as is evidenced by the ability of neural networks of sufficient size to achieve zero training error on image data sets with randomised labels [ZB+21]. The goal of research in this area is both to contribute to the design of better models and provide guarantees and certification for the generalisation performance of models.

In lieu of an in-depth discussion of these measures we refer the reader to the [JN+19], and focus on contextualisation in the reinforcement learning problem and their limitations.

In this section we investigate the offline learning case, where we evaluate a model that the agent has learned offline from limited data. We seek to understand the suboptimality of the expected return $\mathbb{E}_D[V^{\pi^*}(s) - V^{\pi^{D(s)}}(s)]$, using a sample $D \sim D$ as described in 3.2. [FLH+18] decomposes this value further into the sum of the asymptotic bias, and an error due to the size of the dataset D.

$$(V^{\pi^*}(s) - V^{\pi_{D\infty}}(s)) + \mathbb{E}_{D\sim D}[V^{\pi_{D\infty}}(s) - V^{\pi_D}(s)]$$

Due to the very limited direct research on measures for generalisation in reinforcement learning, and enabled by the offline learning setting described above, we largely draw on research on deep learning models. This is an obvious limitation to this review, as to benefit from this section requires the reader to contextualise these methods within their own reinforcement learning models.

One of the primary approaches to studying the generalisation problem in deep learning has been attempting to prove a bound on generalisation gap (ie. the suboptimality of the expected return above). However, so far these bounds have been vacuous [JN+19], and any measures based purely on properties of the model rather than also the data is subject to the issues described in [ZB+21].

An alternative approach is finding complexity measures, ie. properties of the model, that correlate with generalisation performance. This is, subject to the same caveats as empirical measures, where careful experiment design is necessary to indicate causation rather than simple correlation.

Just as sufficiently large neural networks can perfectly fit training data in the supervised learning problem, they can equally do so in the reinforcement learning problem on a simple enough task or simulation environment [ZBP18]. These simple learning problems are the primary means of studying reinforcement learning models, and their drawbacks are discussed in 4.1.

4.3 Verification

As deep reinforcement learning is applied to more safety critical applications, it is important to understand the inputs for which they perform poorly, preferably before those inputs are seen in a real life deployment. Even models that perform well on many examples have been demonstrated to generalise poorly and be susceptible to adversarial settings [WT+11]. We propose that verification of deep learning models is an important avenue of research for interpreting and using generalisation

properties of models in a real world context. To our knowledge, verification of models has not been included in a literature review of generalisation in deep reinforcement learning.

There are a number of different approaches to the verification problem, that may be classified by the approach they use to prove or falsify a model property. Beyond the basic setting of these methods, we omit an in-depth review of verification methods and implementations, and refer the interested reader to [LA+19].

The intent of verifying a system or model is to establish that for an input in a set X, the output of the model belongs to a set Y. Testing the model on a sample of inputs may provide some reassurance as to its behaviour, but due to the size of the input space, which may be infinite in cardinality in some tasks, this does not guarantee the property.

The results of the verification process can be used as evidence to the soundness of the model in a certification process, or to generate adversarial examples to include in the training set to improve the model [KL21a]. For example, DARPA's Assured Autonomy program seeks to use verification of learning models to ensure that the regions of the state space that an agent may explore in the learning process do not overlap with unsafe regions [Nee].

We consider the primary challenges in the application of formal verification of DRL models to be the computational complexity of these methods, and the difficulty of establishing an accurate model of non-trivial environments.

That the scalability of methods for verification of neural networks is a key concern is well established in the literature [KL21b], with current methods falling well short of being able to analyse neural networks of the size being used in industrial applications [BK+20]. For example, the search space generated by a rectified linear unit feed-forward neural network is exponential in the number of its nodes [KL21b] and robustness verification on that network is NP-complete [KB+17]. The complexity of these calculations are further compounded in a sequential decision making problem, such as many of those discussed so far in this paper. Not only must the consequences of the next action be considered, but also the consequences of a arbitrarily long sequence of actions leading up to an unsafe state. Additionally, in a reinforcement learning environment, the effect of the rewards on the model parameters must be considered: as the policy is updated, so is the verification problem.

In order to completely verify a system, it is necessary to have a precise model of the entire system. In the reinforcement learning setting described in 3.2, the system is composed of both the agent and the environment. Therefore in order to understand how the agent could enter an unsafe zone, we must have a perfect model of the environment. Whilst this is possible in games such as chess or go, and perhaps even in settings such as protein folding where there is a set of laws that govern the environment, this is not achievable in most reinforcement learning settings. There have been recent efforts to establish dependability properties on sampled simulation trajectories, which can be transferred to to real environment [DZH21], the performance of the agent on these dependability properties provides another source of measures for generalisation.

5 Conclusion and Directions for Future Research

In this paper we have identified a gap on the literature on generalisation. There is recognition that generalisation is an important issue in AI development, but research so far has focused more on engineering for generalisation rather than establishing a standardised measure for generalisation.

We identified two broad approaches to measuring generalisation: empirical benchmark tests and theoretical methods based on formal properties of models. Both approaches have limitations: empirical benchmarks developed thus far involve simple test environments that are very different from real world environments, and moreover resist scaling up test environment variability. Benchmark tests also do not tell us why a model has generalised well, so good performance on a generalisation benchmark will not necessarily be useful for informing future model design.

For generalisation to be a scientifically rigorous quantity experimental protocols nevertheless need to be developed to measure it. Our first recommendation is that future benchmark tests be designed to scale. By this we mean that the benchmark should be able to scale up both the number of varying environmental parameters, and the extent to which those parameters vary. Real world contexts are complex and change in often unexpected ways. In addition, to combat the issue of overfitting of models by designers to a particular benchmark test, we recommend that a subset of the possible test environments be kept hidden at least for a period of several years. Providing the full range of test environments to research teams from day one would likely lead to models being designed to overfit to the range of test environments. In addition, we recommend that educational tools, be they pre-recorded tutorials or live workshops, should be made available to inform the research community about how to interpret the results of theoretical measures of generalisation and how to determine whether the results of an empirical benchmark tests are applicable to previously unseen environments.

References

[BK+20] Elena Botoeva, Panagiotis Kouvaros, et al. Efficient verification of relu-based neural networks via dependency analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3291–3299, 2020.

[CK+19] Karl Cobbe, Oleg Klimov, et al. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pages 1282–1289. PMLR, 2019.

[DZH21] Yi Dong, Xingyu Zhao, and Xiaowei Huang. Dependability analysis of deep reinforcement learning based robotics and autonomous systems. *arXiv preprint arXiv:2109.06523*, 2021.

[FLH+18] Vincent François-Lavet, Peter Henderson, et al. An introduction to deep reinforcement learning. *arXiv preprint arXiv:1811.12560*, 2018.

[GTR+18] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018.

[HI+18] Peter Henderson, Riashat Islam, et al. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[JN+19] Yiding Jiang, Behnam Neyshabur, et al. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[KB+17] Guy Katz, Clark Barrett, et al. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[KH+21] Brody Kutt, William Hewlett, et al. Innocent until proven guilty (iupg): Building deep learning models with embedded robustness to out-of-distribution content. In *2021 IEEE Security and Privacy Workshops (SPW)*, pages 49–55, 2021.

[KL21a] Panagiotis Kouvaros and Alessio Lomuscio. Towards scalable complete verification of relu neural networks via dependency-based branching. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2643–2650. International Joint Conferences on Artificial Intelligence Organization, 8 2021.

[KL21b] Panagiotis Kouvaros and Alessio Lomuscio. Towards scalable complete verification of relu neural networks via dependency-based branching. In *Proceedings of the 30th international joint conference on artificial intelligence (IJCAI21). To Appear. ijcai. org*, 2021.

[KS+12] Alex Krizhevsky, Ilya Sutskever, et al. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[LA+19] Changliu Liu, Tomer Arnon, et al. Algorithms for verifying deep neural networks. *arXiv preprint arXiv:1903.06758*, 2019.

[LBD+89] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LH+07] Shane Legg, Marcus Hutter, et al. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17, 2007.

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[Nee] Sandeep Neema. Assured autonomy https://www.darpa.mil/program/assured-autonomy.

[PG+18] Charles Packer, Katelyn Gao, et al. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

[SB98] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. 1998.

[SSPS21] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[SSS+17] David Silver, Julian Schrittwieser, Karen Simonyan, et al. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017.

[VBC+19] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019.

[WT+11] Shimon Whiteson, Brian Tanner, et al. Protecting against evaluation overfitting in empirical reinforcement learning. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 120–127, 2011.

[ZB+21] Chiyuan Zhang, Samy Bengio, et al. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[ZBP18] Amy Zhang, Nicolas Ballas, and Joelle Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 2018.

[ZM+21] Chunting Zhou, Xuezhe Ma, et al. Examining and combating spurious features under distribution shift. *arXiv preprint arXiv:2106.07171*, 2021.

[ZV+18] Chiyuan Zhang, Oriol Vinyals, et al. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.

Report

How the group worked together :

- The first thing was to form a team. Students formed an online group where they posted a topic in Computer Science to make a group with other students who were interested in the same. The same went ahead for us, as we formed a group of seven to make a project in the field of Deep Reinforcement Learning. Then it was necessary for us to properly narrow it down to a topic so that everyone can start working on it. We arranged many online and offline meetings to discuss it. Everyone was totally engaged in the process with so many ideas being pitched like weapons control, robotics, gaming, medicine, etc. Finally, Measures of Generalization for Deep Reinforcement Learning was decided as the final topic. We all agreed to choose this because, in real-world situations, the environments are different with many uncertainties and models currently are very limited in terms of generalisability. There has been some recent research in this field of AI, so the idea was also to choose such a topic that is not old but has some literature from where we could start. Also, it was taken into account that the topic could be covered in the given requisite.

- The next task for us was to prepare a presentation on our topic, sort of a preliminary oral report. We divided the tasks for each member. Some found relevant research papers and collected them together to be written as the content for the presentation. A few members who are good at graphics and designing worked on preparing the presentation slides with the information that was being collected. Some members volunteered to present the slides. Overall everyone was involved whether it be reading literature, writing content, formatting of the slides, collecting videos and pictures for it, etc. We created deadlines for each of the tasks and every task that was either to-do or completed was being noted to check the progress of our work. We used various cloud applications such as Google Drive, Google Docs, Google Slides to ensure every piece of literature that was being used, the slides, reports and the whole progress was available to every member of the group, and no work was lost. When the slides were ready, every member of the group reviewed them multiple times to ensure their quality and correct some errors if there were any.

- Before writing the draft of the paper, every member was given more papers to read and collect the relevant material. The information was then discussed and written in the draft. Everyone gave their feedback on how to improve each other's individual work performance. The progress was regularly checked to ensure everyone was on the same page and no one was late for the deadlines we had set. Because of the academic pressure of other modules or health-related reasons, sometimes someone wasn’t able to complete the task in time but someone else always used to step up to finish it which in turn didn’t affect the overall progress of the team. We never felt the need to rush the tasks during the course of the project as the whole group was handling them very responsibly.

- The task to write was divided for each section of the paper. Some worked on writing the introduction, some worked on the abstract, some on the conclusion, and some members who were good at using LaTeX volunteered to write the body while the others supported them by helping summarise the information that was collected in the draft. Everyone also worked on writing this accompanying report which includes how the group worked together, challenges that we faced, solutions to those challenges, and so on.

- Finally, we reviewed all work that was done including the presentation slides, the paper, and the report, in order to find possible revisions that could be done to improve the work and potentially increase the overall score of the team.

Challenges faced:

- Sometimes it was difficult to find the relevant research papers since the topic we are focusing on is new and has limited numbers of good quality literature available up to date.
- Since at the start we did not systematically sort the different types of documents we were collecting in the drive, it was sometimes difficult to find and reference the material. There were many documents including relevant research papers, presentation slides, collection of papers summary, and references.
- Because everyone has got different ideas, it was sometimes difficult for us to be on the same page. For example, when we were choosing a topic to focus on, some wanted a topic related to gaming, some leaned towards robotics, etc. So, we needed to deal with this issue properly because all opinions were worth considering.
- Because of factors such as health, academic pressure the progress of doing individual tasks sometimes got delayed.
- While reading the literature, the material was sometimes difficult to understand and was relatively new for some such as Markov Decision Process, uncertainty in variables, and variability of the environment in Deep Reinforcement Learning.

Solutions to those challenges :

- To find more relevant research papers regarding our topic we searched for more specific keywords such as domain shift, environmental uncertainty, and so on. By doing that, we found enough research papers to work with.
- Google Drive was one of the best solutions to collect all documents in a systematic way. We created sections to categorize the type of works including relevant research papers to record all papers in pdf format, presentation slides, meeting notes to record all important findings from every meeting, and references of all research papers.
- To deal with conflicts, we let each member who proposed the idea explained the main concept, while the others listened and discussed the pros and cons. Then, we decided together on the final decision of the team. If in some cases, we couldn't manage to come to the conclusion, the voting system was implemented. The most voted idea was then the team decision.
- To address the factors of academics and health, whoever got affected by them used to inform the group about the issue and how much time they need as soon as possible so that other members of the team could handle that. The solution was to divide the work smartly among the rest so that the work is not overloaded and deadlines are met.
- In order to understand the material better, some members gained more knowledge about the basic ideas from the internet, books, etc. Some consulted their Undergraduate Professors who

are experts in the Computer Science domain to help them get an understanding of the topic and the field better.

Professional and Ethical Issues

We made sure that we sustain ethical standards regarding plagiarism and properly attribute others' work wherever needed and that we don't replicate others' efforts or ideas.

Due to the nature of a literature review paper, our project did not raise any notable social or ethical issues.

Literature reviews could raise professional issues if the group members were part of an instruction or research team working on the area(s) discussed. The risk would have been that they might overemphasize their or someone else's contribution to the area(s) discussed. This was, however, not applicable in the case of our project as we aren't publishing this paper in an actual journal so the citations won't count for anything.

Conclusion

It was a great learning experience for all of us. We all are from different backgrounds of study and from different countries which made the whole group dynamic and the project very interesting to work on. The project was very new for everyone so there were a lot of learning opportunities which we all made sure to grasp. We encountered a few difficulties here and there but we were able to sort them out smoothly and deliver the requisites on time. Overall, we enjoyed working with each other and on the project and feel our paper will serve as an introduction for people interested in assessing the generalization of their DRL models.