Diverse Auto-Curriculum is Critical for Successful Real-World Multiagent Learning Systems*

Blue Sky Ideas Track

Yaodong Yang†
University College London
Huawei R&D U.K.

Jun Luo
Huawei Canada

Ying Wen
Shanghai Jiao Tong University

Oliver Slumbers
University College London

Daniel Graves
Huawei Canada

Haitham Bou Ammar
Huawei R&D U.K.

Jun Wang
University College London
Huawei R&D U.K.

Matthew E. Taylor
University of Alberta
Alberta Machine Intelligence Institute

ABSTRACT

Multiagent reinforcement learning (MARL) has achieved a remarkable amount of success in solving various types of video games. A cornerstone of this success is the auto-curriculum framework, which shapes the learning process by continually creating new challenging tasks for agents to adapt to, thereby facilitating the acquisition of new skills. In order to extend MARL methods to real-world domains outside of video games, we envision in this blue sky paper that maintaining a diversity-aware auto-curriculum is critical for successful MARL applications. Specifically, we argue that *behavioural diversity* is a pivotal, yet under-explored, component for real-world multiagent learning systems, and that significant work remains in understanding how to design a diversity-aware auto-curriculum. We list four open challenges for auto-curriculum techniques, which we believe deserve more attention from this community. Towards validating our vision, we recommend modelling realistic interactive behaviours in autonomous driving as an important test bed, and recommend the SMARTS benchmark.

KEYWORDS

Multiagent reinforcement learning; auto-curriculum; behaviour models; autonomous driving; simulators; SMARTS

ACM Reference Format:
Yaodong Yang, Jun Luo, Ying Wen, Oliver Slumbers, Daniel Graves, Haitham Bou Ammar, Jun Wang, and Matthew E. Taylor. 2021. Diverse Auto-Curriculum is Critical for Successful Real-World Multiagent Learning Systems: Blue Sky Ideas Track. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 6 pages.

*The authors thank the anonymous reviewers, as well as Greg d'Eon, Calarina Muslimani, Laura Petrich, Sahir, and Amirmohsen Sattarifard for comments and suggestions. Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab, which is supported in parts by research grants from Amii, CIFAR, and NSERC.
†Corresponding author: yaodong.yang@huawei.com

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

1 INTRODUCTION

Reinforcement learning (RL) [85] allows an agent to learn to maximise cumulative rewards via environmental interactions. It has proved successful in many areas, including playing video games [61, 70], robotics control [49], data centre cooling [57], and asset pricing in finance [45]. Multiagent RL (MARL) extends RL to cover the setting where there are multiple learning entities in the environment [36, 97]. This technique has shown remarkable success, especially on multi-player video games such as StarCraft [88], Dota2 [67], and Hide and Seek [5]. However, MARL has had relatively few successes in solving real-world problems. The core thesis of this paper is that the development of learning frameworks that can induce *behavioural diversity* in the policy space is critical for MARL to succeed in real world domains. We summarise existing challenges and recommend autonomous driving (AD) as an ideal test bed for future investigations.

The challenges of deploying RL in the real world are frequently discussed in both workshops [64, 95, 96] and papers [20, 21]. Challenges include the lack of an accurate simulator [13], the high cost of environmental interaction, and the difficulty in learning both effective and diverse policies. Off-policy RL [56] or imitation learning [40] methods could be used for policy evaluation or policy improvement in an offline manner; this would allow agents to learn a good initial behaviour before ever interacting with an environment. However, these methods are only applicable if training data sets exist. When the training data are limited, for example in the AD domain [100], offline methods are insufficient for robust performance in the real world due to a lack of *diversity* in agents' behaviours [18, 92]. In fact, even in cases when a simulator is available, lack of behavioural diversity could still exist due to the sim-to-real gap [59, 65, 76]. Unfortunately, MARL suffers from all of these concerns. Furthermore, the additional complexity of multi-agent problems that arises from the cross product of multiple agents' state and action spaces induced by social interactions compounds these concerns. Developing frameworks that can deal with the underlying complexities of the MARL domain is crucial. We argue

that the development of effective, yet diverse behaviours is critical for MARL to have an impact in domains outside of video games.

The *auto-curricula framework* [53, 72] is a promising direction towards such a goal. In natural evolution, species with stronger adaptability flourish when nature alters the environment and some previously well-adapted species no longer survive. Through this *co-evolution process* [22, 68, 74], the diversity of life on Earth has been maintained over billions of years. Inspired by this mechanism of bio-diversity in nature, a series of MARL learning frameworks have recently been proposed and have demonstrated remarkable empirical success. These include open-ended evolution [9, 50, 82], population based training [42, 58], and training by emergent curricula [5, 53, 72]. In general, these frameworks can be unified under the idea of an auto-curriculum where an endless procession of better-performing agents are automatically generated by exerting selection pressure among the multiple self-optimising agents. The underlying principle of auto-curricula is that any adaptation an agent makes will have a cascading effect that other agents must adapt to in order to survive. This intrinsically provides that provides an automatic curriculum that continually facilitates agents' acquisition of new skills via such social interactions.

In this Blue Sky paper, we emphasise that maintaining a diversity-aware auto-curriculum is critical for successful MARL applications. Specifically, we advocate verification through one specific real-world problem: modelling interactive behaviours in AD scenarios. The main contributions of this work are as follows. We start by highlighting the necessity of behavioural diversity in multiagent systems in Section 2 and then briefly survey existing works and explain why they are not yet sufficient for MARL in Section 3. We investigate the idea of auto-curriculum in Section 4, and raise four open challenges which we believe deserve more attention from this community. Section 5 discusses why AD is an excellent domain to host such investigations and proposes one particular test bed for future study. Finally, we reiterate our vision in Section 6.

2 THE NECESSITY OF DIVERSITY

Nature exhibits a remarkable tendency towards *diversity* [38]. Over the past billions of years, a vast assortment of unique species have naturally evolved. Each one is capable of orchestrating the complex biological processes necessary to sustain life. Analogously, in computer science, machine intelligence can be considered as the ability to adapt to a diverse set of complex environments [37]. This suggests that the ceiling of intelligence rises when environments of increasing diversity and complexity are provided. In fact, recent successes in developing AI capable of achieving super-human performance on complicated multi-player video games, such as StarCraft [34, 89], Honour of King [99], Hide and Seek [5], and Dota2 [67], have provided justification for emphasising behavioural diversity when designing learning protocols in multiagent systems. Specifically, promoting behavioural diversity is pivotal for MARL methods. Diversity not only prevents AI agents from checking the same policies repeatedly, but also helps agents discover niche skills, avoid systematic weaknesses and maintain robust performance when encountering unfamiliar types of opponents at test time.

Behavioural diversity and the *non-transitivity* of many environments are intertwined. In biological systems, bio-diversity is promoted by the non-transitive interactions among many competing populations [43, 75]. The central feature of such non-transitive relations can be thought of as analogous to a Rock-Paper-Scissors game, where rock beats scissors, scissors beats paper, and paper beats rock. In game theory, the necessity of pursuing behavioural diversity is also deeply rooted in the non-transitive structure of games [6, 8, 48]. In general, an arbitrary game, of either the normal-form type [15] or the differential type [7], can always be decomposed into a sum of two components: a *transitive part* plus a *non-transitive part*. The transitive part of a game represents the structure in which the rule of winning is transitive (i.e., if strategy A beats B, B beats C, then A can surely beat C), and the non-transitive part refers to the game structure in which the set of strategies follow a cyclic rule (e.g., the endless cycles of rock, paper, and scissors). Diversity matters especially for the non-transitive part because there is no consistent winner in such sub-games: if a player only plays rock, he can be exploited by paper, but not so if he is diverse in playing pock and scissors. In fact, real-world problems often consist of a mixture of both parts [17], therefore it is critical to design learning objectives that lead to behavioural diversity.

Effective MARL performance often requires diversity in two aspects. The first aspect is about the training player uses diversified strategies against a fixed type of opponents. Most games involve non-transitivity in the policy space and thus it is necessary for each player to acquire a diverse set of winning strategies to achieve high unexploitability. The second aspect is the ability to pick a diverse set of opponents/teammates during training. In playing cooperative card games like Hanabi [10], one player may or may not understand the indirect signalling when choosing a card to play. If an agent has not learned to play diversely under both mindsets, it will fail to accurately model the collaborator and play sub-optimally. Similarly, in real-world driving, distinct locales have different conventions. The UK and the US drive on different sides of the road, or even within the same country, different cities can follow different conventions. For example, the *Pittsburgh Left* convention assumes that a few cars will turn left in front of traffic at the beginning of a green light [94], while other areas assume that cars will turn left in front of traffic during a yellow or at the beginning of a red light [78]. As a result, it can be expected that an autonomous agent without a diverse mindset could create hazards on the road [93].

3 RELATED WORK ON DIVERSITY

Despite the importance of diversity, there has been limited work within the machine learning domain where diversity is modelled in a principled way. Furthermore, there is no agreed upon, formal definition. For example, behavioural diversity can be defined as the variance in rewards [50, 51], the convex hull of a *gamescape* [6], choosing whether or not to visit a new environmental state [27, 84, 98], or, acquiring new types of skills in a task [28, 35].

So far, the majority of work that models diversity lies in evolutionary computation (EC) [3, 29], which attempts to mimic the natural evolution process. One classic idea in EC is *novelty search* [50, 51], which aims to search for behaviours that lead to different outcomes. Quality-diversity (QD) methods hybridise novelty search with fitness under the notion of survival of the fittest (i.e., high utility) [73]. Two representatives are *Novelty Search with Local Competition* [52] and *MAP-Elites* [16, 62].

Searching for behavioural diversity is also a common topic in RL, which is often studied in the context of skill discovery [27, 28, 35], intrinsic rewards [11, 12, 31], or maximum-entropy learning [32, 33, 55]. These RL algorithms can be considered as QD methods, in the sense that quality refers to maximising cumulative reward, and diversity means either visiting a new state [27, 98] or obtaining a policy with larger entropy [55].

In the context of MARL, learning typically means an agent acting in an open-ended system with continually changing policies by different opponents. Yet, such a learning process can only guarantee *differences* but not *diversity*, which are two different notions–diversity is not an inherent feature in MARL. In fact, understanding the principle of how diversity is promoted in an auto-curriculum is an open problem in MARL [6, 69, 98]. In the example of training soccer AIs [47], learning against only *different* opponents can easily make an agent get into circular dynamics and not improve. Finally, this work is also different from the previous manifesto [53], which links the auto-curricula in natural evolution with MARL; our main focus is to emphasise creating diverse auto-curricula.

4 EXISTING OPEN CHALLENGES

Auto-curricula [5, 53, 72] provide a framework to automatically shape learning procedures for AI agents by consistently challenging them with new tasks that are adapting to their capabilities. As the challenges generated by an auto-curriculum become increasingly diverse and complex over time, AI agents accumulate more diverse and effective skills. In fact, recent successes in training AIs that achieve super-human performance and acquire diverse behaviours on complex video games [5, 80, 89] provide strong justification for adopting auto-curricula a diversifying learning protocol. However, in order to serve as a general framework to tackle more real-world problems beyond video games, auto-curriculum technique still faces four open challenges.

Open Challenges of Designing Diversity-Aware Auto-Curricula

(1) How do we measure diversity in an auto-curriculum?
(2) How do we generate diversity-aware auto-curricula, especially in non-zero sum settings?
(3) How do we shape an auto-curriculum to induce diverse yet effective behaviours?
(4) How do we deal with non-transitivity when learning in a diversity-aware auto-curricula?

The first challenge is to define the correct objective to measure and promote diversity in the generated auto-curriculum. In the single-agent setting, diversity can be defined through a different reward function [50, 51], visiting a new state [27, 84, 98], or acquiring a new skill [28, 35]. However, in the MARL setting, with multiple players, each having a population of strategies, diversity should be defined in the joint policy space, considering all existing strategies of all agents. Yet, there is limited work that tries to quantify the behavioural diversity at the population level. Although there are no straightforward answers, we believe one promising direction could be to leverage the *determinantal point process* [46] from quantum physics. This processs measuring diversity through the determinant value in a vector space, thus the level of orthogonality among the input vectors can be represented by agents' different joint-strategy profiles in terms of rewards [98].

The second challenge involves the applicability on *non-zero sum games*. The curricula in the examples of StarCraft [88] or Hide and Seek [5] are generated by competitive self-play from the players in zero-sum games. However, many real-world tasks, such as autonomous driving, are not zero-sum—in fact, they tend to be a mixed setting where cooperation outweighs competition. Therefore, creating an auto-curriculum in non-zero sum games is an open problem. Interestingly, recent studies have shown that adapting in social dilemmas can create an effective auto-curriculum for the emergence of collective cooperation [39, 54, 71]. Through sequences of new challenges in addressing social dilemmas, agents eventually learn to achieve a socially-beneficial outcome. This resembles tasks such as discovering collective driving strategies that can mitigate congestion. For example, consider the case of solving Braess's paradox [14] (i.e., a typical example in modelling road network and traffic flow) where agents progressively learn to sanction those who tend to over-exploit the common resources, thus creating new curriculum. Nonetheless, creating curricula for collective cooperation is still under-developed relative to auto-curricula induced by zero-sum games. Importantly, as pointed out by Leibo et al. [53], auto-curricula induced by social dilemmas could be cursed by the "no-free-lunch" property: once you resolve a social dilemma in one place, another one crops up to take its place, a problem also known as higher-order social dilemmas [60, 66].

Thirdly, although RL techniques offer insight into how a desirable behaviour can be learned in a fixed environment, it is still unclear how complex and useful behaviours can be best developed, while these behaviours are influencing the environment. In fact, it is often the case that the more complex the behaviour, the less likely it is generated completely from scratch [53]. An example is that it is highly unlikely a world-champion level policy is quickly generated by a curriculum when learning to act in complex environments. Moreover, this issue is only exacerbated when multiple agents ($N \gg 2$) are involved to explore the joint-strategy space. Fortunately, initial progress has been made by works on *Policy Space Response Oracle (PSRO)* [6, 48, 63] where different kinds of *rectifiers* have been proposed to shape the auto-curricula so that effective behaviours with high quality can be emphasised. For example, PSRO with a *Nash rectifier* [6] explores only strategies that have positive Nash support so as to preserve the strategy strength. Despite the empirical success in generating diverse yet effective strategies, PSRO methods only work in solving symmetric zero-sum games, a limitation highlighted in Open Challenge II.

Lastly, results on game decomposition suggests that a game [6, 15] generally consists of both *transitive* and *non-transitive* structures. The topological structure of real-world tasks often resembles a spinning top if projected onto a 2D space [17], where the x-axis is the non-transitive dimension and y-axis is the transitive dimension. The non-transitive part can harm the effectiveness of auto-curriculum [6, 17]. For example, an auto-curriculum generated by self-play in zero-sum games [30, 77] could make a learning agent endlessly chase its own tail by creating the same tasks repetitively without breaking out. Things become even worse when the non-transitivity issue couples with the catastrophic-forgetting property

of the model itself, as seen with deep neural networks [44]. As a result, agents may end up with acquiring mediocre solutions or getting trapped in limited cycles within the strategy space [6, 8]. Memorising a library of all possible policies can help prevent cycling (e.g., three strategies in the toy example of Rock-Paper-Scissors), but for many real-world tasks, the dimension of the non-transitive cycles can be huge, and as a result, building such a library itself becomes an endless task [17].

5 AN AUTONOMOUS DRIVING TEST BED

Creating a diversity-aware auto-curricula for MARL, although still facing several open challenges, is a critical step for deploying successful multi-agent learning systems in real-world domains. For validating effectiveness, we believe autonomous driving (AD) simulation environments provide an excellent test bed.

AD technologies [4] enable a vehicle to sense its environment and move safely to a destination with little or no human intervention. Since the first DARPA competition in 2004, where the best performing car completed only 7.3 miles of the 142-mile desert route, remarkable progress has been made. For example, the commercial company Waymo has driven more than 20 million miles on public roads under the SAE level-4 setting [41, 93].

In spite of such achievements, fully competent and natural interactions with other road users remain out-of-reach. Rather than embracing inter-driver interaction, current mainstream level-4 AD solutions restricts it. When encountering complex interactive scenarios, autonomous cars tend to slow down and wait for the situation to become more simple. They rarely cut in front of another car or force its way in at a merge, as human drivers routinely do. In California in 2018, 86% of crashes involving autonomous vehicles were attributable to the AD car's conservative behaviour [83], with 57% rear endings and 29% sideswipes by other vehicles on the AD car. Trial AD cars in Arizona and California are often targets of complaints for blocking other cars [79], excessive hard braking [23], hesitant highway merging, and inflexible pick-up/drop-off locations [24, 25]. While rarely illegal, the overly conservative driving style frustrates human drivers, and can even pose road hazards. This also restricts AD technologies from being applied on special-purpose vehicles, such as ambulances or police cars, where aggressive driving behaviour may be required.

A key reason for this limitation is that existing AD *simulators* have limited capacity for modelling realistic interactions with diverse driving behaviours. Simulators are crucial for validation of the AI software controlling the autonomous vehicle (also called *ego vehicle*). For validating the ego vehicle's interactive behaviour with *social vehicles* (i.e., other vehicles that share the same driving environment), we need diverse *social agents* capable of realistic and competent interaction. Conventional AD simulators [19, 81, 90] focus on modelling sensory inputs and control dynamics, rather than interaction. As a result, the behaviours of social vehicles end up being controlled by simple scripts or rule-based models (e.g., IDM for longitudinal control [86] or MOBIL for lateral control [87]) and the simulated interaction between ego and social vehicles falls short of the richness and diversity seen in the real world. AD companies heavily use replay of historical data collected from real-world trials to validate ego vehicle behaviour [2]. However, such a data-replay approach to simulation does not allow true interaction between vehicles because the social vehicles are not controlled by intelligent agents but merely stick to historical trajectories [1, 91]. In short, how to create a population of diverse intelligent social agents that can be adopted in simulation to provide traffic with realistic interaction is still an open question.

We believe that a MARL approach powered by diversity-aware auto-curriculum can help solve the problem of generating high quality behaviour models that approach human-level sophistication. Fundamentally, driving in a shared public space with other road users is a multi-agent problem wherein the behaviour of agents co-evolve. Co-evolved diverse and competent behaviours can allow AD simulation to encompass sophisticated interactions seen among human drivers and thus alleviate the conservativeness of existing AD solutions. Crucially, solving this problem for AD will also require the key research challenges identified in Section 4 to be addressed.

For such a plan to work, we need an appropriate simulator that supports MARL auto-curriculum for diversity, such as the SMARTS AD simulator [101]. Unlike other existing simulators, SMARTS is *natively multi-agent*. Social agents use the same APIs as the ego agent to control vehicles, and thus may use arbitrarily complex computational models, either rule-based or (MA)RL-driven. The SMARTS *social agent zoo* hosts a growing number of behaviour models to be used by simulated agents, regardless of their divergence in model architectures, observation/action spaces, or computational requirements. The key component that makes such computations possible is the built-in "bubble" mechanism. This mechanism defines a spatiotemporal region so that intensive computing is only activated inside the bubbles where fine-grained interaction is required, such as at unprotected left-turns, roundabouts, and highway double merges. These features make SMARTS highly suitable for multi-agent auto-curriculum studies.

To be more concrete, consider the ULTRA [26] benchmark suite built on top of SMARTS. It includes over 100,000 unprotected left-turn scenarios at different levels of complexity. These scenarios could seed an auto-curriculum that gradually increases the diversity and complexity of interaction by injecting trained behaviour models through available SMARTS mechanisms, allowing for not only more diverse agent behaviour, but also a curriculum composed of increasingly difficult scenarios. Through this, we expect interaction behaviours reminiscent of the *Pittsburgh Left* to emerge in a fashion that could potentially reach the level of sophistication of human drivers. In turn, such emergent behavioural models can be used to support diverse interactions in AD simulations far beyond what has been possible through the rule-based models.

6 CONCLUSION

Despite remarkable success shown on video games, we envision that developing a diversity-aware auto-curriculum framework is a critical step to ensure the success of MARL technique on real-world problems. Specifically, we believe the pressing need for high-quality behaviour models in AD simulation is a great opportunity for the MARL community to make a unique contribution by (1) theoretically addressing the modelling challenges on behavioural diversity and (2) experimentally training generations of increasingly diverse agents to provide interactions for real-world AD.

REFERENCES

[1] Drago Anguelov. 2020. Presentation at MIT course on Self-Driving Cars. https://www.youtube.com/watch?v=Q0nGo2-y0xY&feature=youtu.be&t=1920&ab_channel=LexFridman.

[2] The Atlantic. 2017. Inside Waymo's Secret World for Training Self-Driving Cars. https://www.theatlantic.com/technology/archive/2017/08/inside-waymos-secret-testing-and-simulation-facilities/537648/.

[3] Thomas Bäck, David B Fogel, and Zbigniew Michalewicz. 1997. Handbook of evolutionary computation. *Release* 97, 1 (1997), B1.

[4] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius Brito Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago Meireles Paixão, Filipe Mutz, et al. 2020. Self-driving cars: A survey. *Expert Systems with Applications* (2020), 113816.

[5] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528* (2019).

[6] D Balduzzi, M Garnelo, Y Bachrach, W Czarnecki, J Pérolat, M Jaderberg, and T Graepel. 2019. Open-ended learning in symmetric zero-sum games. In *ICML*, Vol. 97. PMLR, 434–443.

[7] D Balduzzi, S Racaniere, J Martens, J Foerster, K Tuyls, and T Graepel. 2018. The Mechanics of n-Player Differentiable Games. In *ICML*, Vol. 80. JMLR. org, 363–372.

[8] David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. 2018. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems*. 3268–3279.

[9] Wolfgang Banzhaf, Bert Baumgaertner, Guillaume Beslon, René Doursat, James A Foster, Barry McMullin, Vinicius Veloso De Melo, Thomas Miconi, Lee Spector, Susan Stepney, et al. 2016. Defining and simulating open-ended novelty: requirements, guidelines, and challenges. *Theory in Biosciences* 135, 3 (2016), 131–161.

[10] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[11] Andrew G Barto. 2013. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*. Springer, 17–47.

[12] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*. 1471–1479.

[13] Joschka Boedecker and Minoru Asada. 2008. Simspark–concepts and application in the robocup 3d soccer simulation league. *Autonomous Robots* 174 (2008), 181.

[14] Dietrich Braess. 1968. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12, 1 (1968), 258–268.

[15] Ozan Candogan, Ishai Menache, Asuman Ozdaglar, and Pablo A Parrilo. 2011. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research* 36, 3 (2011), 474–503.

[16] Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. 2015. Robots that can adapt like animals. *Nature* 521, 7553 (2015), 503–507.

[17] Wojciech Marian Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. 2020. Real World Games Look Like Spinning Tops. *arXiv* (2020), arXiv–2004.

[18] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. 2019. RoboNet: Large-Scale Multi-Robot Learning. *arXiv* (2019), arXiv–1910.

[19] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. *arXiv preprint arXiv:1711.03938* (2017).

[20] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. 2020. An empirical investigation of the challenges of real-world reinforcement learning. *arXiv preprint arXiv:2003.11881* (2020).

[21] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. 2019. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901* (2019).

[22] William H Durham. 1991. *Coevolution: Genes, culture, and human diversity*. Stanford University Press.

[23] Amir Efrati. 2020. Waymo Riders Describe Experiences on the Road. https://www.theinformation.com/articles/waymo-riders-describe-experiences-on-the-road

[24] Amir Efrati. 2020. Waymo's Backseat Drivers: Confidential Data Reveals Self-Driving Taxi Hurdles. https://www.theinformation.com/articles/waymos-backseat-drivers-confidential-data-reveals-self-driving-taxi-hurdles

[25] Amir Efrati. 2020. Waymo's Big Ambitions Slowed by Tech Trouble. https://bit.ly/31lKwgt.

[26] Mohamed Elsayed, Kimia Hassanzadeh, Nhat M. Nguyen, Montgomery Alban, Xiru Zhu, Daniel Graves, and Jun Luo. 2020. ULTRA: A reinforcement learning generalization benchmark for autonomous driving.

[27] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is All You Need: Learning Skills without a Reward Function. In *International Conference on Learning Representations*.

[28] Carlos Florensa, Yan Duan, and Pieter Abbeel. 2017. Stochastic Neural Networks for Hierarchical Reinforcement Learning. *arXiv* (2017), arXiv–1704.

[29] David B Fogel. 2006. *Evolutionary computation: toward a new philosophy of machine intelligence*. Vol. 1. John Wiley & Sons.

[30] Michael E Gilpin. 1975. Limit cycles in competition communities. *The American Naturalist* 109, 965 (1975), 51–60.

[31] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. 2017. Variational Intrinsic Control. (2017).

[32] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement Learning with Deep Energy-Based Policies. In *ICML*.

[33] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*. 1861–1870.

[34] Lei Han, Jiechao Xiong, Peng Sun, Xinghai Sun, Meng Fang, Qingwei Guo, Qiaobo Chen, Tengfei Shi, Hongsheng Yu, and Zhengyou Zhang. 2020. TStarBot-X: An Open-Sourced and Comprehensive Study for Efficient League Training in StarCraft II Full Game. *arXiv preprint arXiv:2011.13729* (2020).

[35] Karol Hausman, Jost Tobias Springenberg, Ziyu Wang, Nicolas Heess, and Martin Riedmiller. 2018. Learning an embedding space for transferable robot skills. In *International Conference on Learning Representations*.

[36] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 750–797.

[37] José Hernández-Orallo. 2017. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press.

[38] John Henry Holland et al. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

[39] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*. 3326–3336.

[40] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–35.

[41] SAE International. 2014. Automated Driving Levels of Driving Automation are Defined in New SAE International Standard J3016.

[42] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 364, 6443 (2019), 859–865.

[43] Benjamin Kerr, Margaret A Riley, Marcus W Feldman, and Brendan JM Bohannan. 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature* 418, 6894 (2002), 171–174.

[44] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[45] Petter N Kolm and Gordon Ritter. 2020. Modern perspectives on reinforcement learning in finance. *Modern Perspectives on Reinforcement Learning in Finance (September 6, 2019). The Journal of Machine Learning in Finance* 1, 1 (2020).

[46] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083* (2012).

[47] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4501–4510.

[48] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in neural information processing systems*. 4190–4203.

[49] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. 2020. Learning quadrupedal locomotion over challenging terrain. *Science robotics* 5, 47 (2020).

[50] Joel Lehman and Kenneth O Stanley. 2008. Exploiting open-endedness to solve problems through the search for novelty.. In *ALIFE*. 329–336.

[51] Joel Lehman and Kenneth O Stanley. 2011. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation* 19, 2 (2011), 189–223.

[52] Joel Lehman and Kenneth O Stanley. 2011. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th*

annual conference on Genetic and evolutionary computation. 211–218.

[53] Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. 2019. Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research. *arXiv* (2019), arXiv–1903.

[54] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 464–473.

[55] Sergey Levine. 2018. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909* (2018).

[56] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[57] Yuanlong Li, Yonggang Wen, Dacheng Tao, and Kyle Guan. 2019. Transforming cooling optimization for green data center via deep reinforcement learning. *IEEE transactions on cybernetics* 50, 5 (2019), 2002–2013.

[58] Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, and Thore Graepel. 2018. Emergent Coordination Through Competition. In *International Conference on Learning Representations*.

[59] Jan Matas, Stephen James, and Andrew J Davison. 2018. Sim-to-real reinforcement learning for deformable object manipulation. *arXiv preprint arXiv:1806.07851* (2018).

[60] Sarah Mathew. 2017. How the second-order free rider problem is solved in a small-scale society. *American Economic Review* 107, 5 (2017), 578–81.

[61] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[62] Jean-Baptiste Mouret and Jeff Clune. 2015. Illuminating search spaces by mapping elites. *arXiv* (2015), arXiv–1504.

[63] Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, et al. 2019. A Generalized Training Approach for Multiagent Learning. In *International Conference on Learning Representations*.

[64] Workshop of RL4RealLife. 2020. Rl for Real Life 2020. https://sites.google.com/view/RL4RealLife.

[65] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. 2018. Learning Dexterous In-Hand Manipulation. *CoRR* abs/1808.00177 (2018). arXiv:1808.00177 http://arxiv.org/abs/1808.00177

[66] Elinor Ostrom. 2000. Collective action and the evolution of social norms. *Journal of economic perspectives* 14, 3 (2000), 137–158.

[67] Jakub Pachocki, Greg Brockman, Jonathan Raiman, Susan Zhang, Henrique Pondé, Jie Tang, Filip Wolski, Christy Dennison, Rafal Jozefowicz, Przemyslaw Debiak, et al. 2018. OpenAI Five, 2018. *URL https://blog. openai. com/openai-five* (2018).

[68] Jan Paredis. 1995. Coevolutionary computation. *Artificial life* 2, 4 (1995), 355–375.

[69] Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, and Stephen Roberts. 2020. Effective Diversity in Population-Based Reinforcement Learning. *arXiv preprint arXiv:2002.00632* (2020).

[70] Peng Peng, Ying Wen, Yaodong Yang, Quan Yuan, Zhenkun Tang, Haitao Long, and Jun Wang. 2017. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069* (2017).

[71] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*. 3643–3652.

[72] Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. 2020. Automatic Curriculum Learning For Deep RL: A Short Survey. *arXiv preprint arXiv:2003.04664* (2020).

[73] Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. 2016. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI* 3 (2016), 40.

[74] Mark D Rausher. 2001. Co-evolution and plant resistance to natural enemies. *Nature* 411, 6839 (2001), 857–864.

[75] Tobias Reichenbach, Mauro Mobilia, and Erwin Frey. 2007. Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. *Nature* 448, 7157 (2007), 1046–1049.

[76] Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. 2017. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot Learning*. PMLR, 262–270.

[77] Spyridon Samothrakis, Simon Lucas, ThomasPhilip Runarsson, and David Robles. 2012. Coevolving game-playing agents: Measuring performance and intransitivities. *IEEE Transactions on Evolutionary Computation* 17, 2 (2012), 213–226.

[78] Valley Driving School. 2018. 6 Times You Can Proceed on a Red Light. https://www.valleydrivingschool.com/blog/main/6-times-you-can-proceed-on-a-red-light.

[79] Faiz Siddiqui. 2020. Some of the biggest critics of Waymo and other self-driving cars are the Silicon Valley residents who know how they work. https://wapo.st/30UAX6R.

[80] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[81] LGSVL Simulator: An Autonomous Vehicle Simulator. 2020. LGSVL Simulator. https://www.lgsvlsimulator.com/docs/.

[82] Russell K Standish. 2003. Open-ended artificial evolution. *International Journal of Computational Intelligence and Applications* 3, 02 (2003), 167–175.

[83] Jack Stewart. 2020. Humans Just Can't Stop Rear-Ending Self-Driving Cars – Let's Figure Out Why. https://bit.ly/3jfcfFs.

[84] Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O Stanley, and Jeff Clune. 2017. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv preprint arXiv:1712.06567* (2017).

[85] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[86] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. 2000. Congested traffic states in empirical observations and microscopic simulations. *Physical review E* 62, 2 (2000), 1805.

[87] Martin Treiber and Arne Kesting. 2009. Modeling lane-changing decisions with MOBIL. In *Traffic and Granular Flow'07*. Springer, 211–221.

[88] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[89] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[90] Virtual Test Drive (VTD). 2020. Complete Tool-Chain for Driving Simulation. https://www.mscsoftware.com/product/virtual-test-drive.

[91] Kyle Vogt. 2020. The Disengagement Myth. https://medium.com/cruise/the-disengagement-myth-1b5cbdf8e239.

[92] Ziyu Wang, Josh S Merel, Scott E Reed, Nando de Freitas, Gregory Wayne, and Nicolas Heess. 2017. Robust imitation of diverse behaviors. *Advances in Neural Information Processing Systems* 30 (2017), 5320–5329.

[93] Waymo. 2020. Waymo Safety Report. https://bit.ly/2T4vRl4.

[94] Wikipedia. 2020. Pittsburgh Left. https://en.wikipedia.org/wiki/Pittsburgh_left.

[95] ICML Workshop. 2019. Reinforcement Learning for Real Life. https://icml.cc/Conferences/2019/ScheduleMultitrack?event=3515.

[96] NeurIPS Workshop. 2020. Challenges of Real-World RL. https://sites.google.com/view/neurips2020rwrl.

[97] Yaodong Yang and Jun Wang. 2020. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *arXiv preprint arXiv:2011.00583* (2020).

[98] Yaodong Yang, Ying Wen, Lihuan Chen, Jun Wang, Kun Shao, David Mguni, and Weinan Zhang. 2020. Multi-Agent Determinantal Q-Learning. (2020).

[99] Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, et al. 2020. Towards Playing Full MOBA Games with Deep Reinforcement Learning. *arXiv e-prints* (2020), arXiv–2011.

[100] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. 2019. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088* (2019).

[101] Ming Zhou, Jun Luo, Julian Villela, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. 2020. SMARTS: Scalable Multi-Agent Reinforcement Learning Training School for Autonomous Driving. *arXiv preprint arXiv:2010.09776* (2020).