Constrained Reinforcement Learning for Autonomous Farming: Challenges and Opportunities

Yongshuai Liu, Taeyeong Choi, and Xin Liu

University of California, Davis
yshliu@ucdavis.edu, taechoi@ucdavis.edu, xinliu@ucdavis.edu

Abstract

Reinforcement learning (RL) is a machine learning technique to optimize policies to execute the most desirable sequence of actions in a complex environment. In particular, the optimal set of parameters is automatically learned based on observations while continuously interacting with environments. While this powerful approach has shown great success in various applications (e.g. games, robotics, etc.), relatively little attention has been paid to the agriculture domain. In this paper, we first discuss a general framework for RL with constraint terms for agricultural scenarios and also explore the potential challenges in developing it into a successful model under realistic environments. We then introduce potential data-driven strategies to effectively mitigate those challenges to realize fully autonomous systems for farm management.

1 Introduction

With the rapid growth of global population, the demand for healthy and fresh food is rising fast (FAO 2022). To improve sustainability of current food systems, novel AI and machine learning techniques have been proposed for precision agriculture, where an individual-specific treatment is applied to each crop not only to maximize the total yields but also to utilize the available resources in an efficient manner (Roopaei, Rad, and Choo 2017). For such a decision-making process, reinforcement learning (RL)—shown to be successful in various applications, such as games (Vinyals et al. 2019), robotics (Agostinelli et al. 2019), and recommender systems (Afsar, Crump, and Far 2021)—has also been studied to learn optimal policies in agricultural domains (Cao et al. 2022; Ajagekar and You 2022).

In this paper, we consider the agricultural control problem in which an RL controller can access a variety of sensors and actuators to estimate the states of the whole system as well as drive it to a desirable condition. To be specific, Fig. 1 illustrates the closed-loop framework of RL for a greenhouse that can repeatedly determine the parameters of built-in actuators to adjust physical properties (e.g., temperature, humidity, and lighting) based on the latest sensor readings of the climate and plants. Moreover, the RL controller would continuously improve the policy based on the crop status.

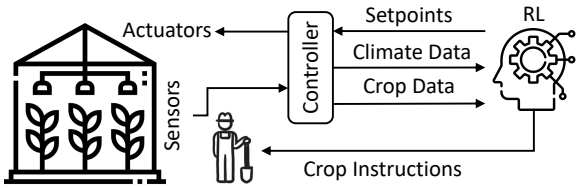

Figure 1: Illustration of the control framework.

Similarly, some specialized sensors could be deployed to automatically decide to gate in chemical fertilizers or water to manage the soil when necessary. The ultimate objective of this autonomous system is to precisely maintain the best environmental conditions to maximize the productivity of crop while controlling the cost of such resources.

Recently, Gandhi (2022) and Gautron et al. (2022) have surveyed some state-of-the-art technologies for such applications. This paper, however, offers novel perspectives that have not been covered albeit essential for real use cases:

- In Sec. 2, we formulate the decision-making process of RL into a stochastic Constrained Markov Decision Process (CMDP) instead of a naïve MDP to directly involve realistic resource constraints in optimization.
- Practical challenges (cf. Sec. 3) are each linked to specific potential opportunities (cf. Sec. 4) that would be well incorporated into the general pipeline of RL for agricultural applications.

2 Constrained Reinforcement Learning

Constrained RL aims to train a *controller* to perform appropriate *actions* based on the observed *states* to maximize the total *reward* under *cost* constraints. Here, we suggest formulating the sequential decision-making process of autonomous farming as a stochastic Constrained Markov Decision Process (CMDP) (Liu et al. 2021), which is a constrained optimization problem in essence. The objectives can be set to maximize the quantities positively correlated to the crop yield (e.g., volume, weight, growth rate, etc.) or minimize the occurrence of negative events (e.g., unhealthy fruits). Concurrently, the farming controller may be subject to a number of constraints on usable resources—e.g., water, electricity, and fertilizers—as well as safety requirements, such as minimal level of temperature.

Formally, we represent the farming control problem as a CMDP with tuple (S, A, R, C, P, γ). To be specific, the observations from deployed sensors (i.e., temperature, images of plant, etc.) at each time t constitute the state $s_t \in S$ of the whole system. In addition, the action set A consists of possible joint actions a_t (e.g., open air inlet) for accessible actuators (e.g., ventilation system) to influence certain properties (e.g., air quality). Also, the reward function $R : S \times A \times S \mapsto \mathbb{R}$ and the cost function $C : S \times A \times S \mapsto \mathbb{R}$ are both dependent upon the taken action a_t under state s_t and its resulting sate s_{t+1}. Multiple cost functions can be set up so that each can correspond to a particular constraint. Moreover, the state transition function $P : S \times A \times S \mapsto [0, 1]$ represents $\Pr(s_{t+1}|s_t, a_t)$, which governs the stochastic dynamics of the world, albeit unknown in most realistic scenarios. Lastly, γ is the discount factor, possibly defined to be different for reward and cost calculations.

In particular, as suggested in (Liu, Halev, and Liu 2021), we also consider two types of constraints: 1) *Cumulative* constraints require the accumulated costs to be within a certain limit (e.g., the total usage of water) and 2) *Instantaneous* ones require each action/state not to violate any pre-defined condition at each time slot—e.g., minimal humidity.

Eventually, RL is designed to learn a policy π that takes state s_t as input to output an action a_t, which leads to state s_{t+1}, determined by the state transition function P. In particular, the optimal policy π^* is to maximize the discounted cumulative reward $J_R^\pi = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$, while satisfying both discounted cumulative constraints $J_{C_i}^\pi = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t, s_{t+1})]$ and instantaneous constraints C_j, where $\tau = (s_0, a_0, s_1, a_1...)$ is a trajectory. Thus, the final form for optimization can be defined as:

$$\underset{\pi}{\text{maximize}} \quad \max_\pi J_R^\pi \tag{1}$$

$$\text{subject to} \quad J_{C_i}^\pi \leq \omega_i, \text{for each i}, \tag{2}$$

$$C_j(s_t, a_t) \leq \epsilon_j, \text{for each j and t}. \tag{3}$$

3 Practical Challenges

In this section, we describe the technical challenges to apply RL in agricultural domains. Even though some of the challenges may also exist in other domains, it is more severe in agriculture systems because of the domain-specific natures.

3.1 System Complexity

Farming is by nature a highly complex activity since it requires to model various environmental factors as well as the consequences from the interactions among them in plants, soils, and fertilizers over a long period of time. Furthermore, the same species of crop can grow into diverse *phenotypes* even under a carefully controlled condition, which would promote different interactions with environments. As a result, the state-action space for RL may be prohibitively large for a controller to evaluate every possible state-action pair during training. Trained controllers would, thus, have to keep performing interpolation or extrapolation over unseen inputs after deployment, so taken actions may not be reliable—i.e., poor generalizability (Lee et al. 2019). Moreover, accurate physics-based simulators cannot be easily built as interactive testbeds for RL, which also hinders active research.

3.2 Resource Constraints & Safety Requirements

The primary goal of smart farming is not only to improve productivity but to *restrict* the used resources—such as water and electricity usage—to a certain limit. Moreover, as in other RL problems with physical environments, learning cannot be performed by freely exploring possible actions with trail and error, because otherwise some undesirable, *irreversible* outcomes (e.g., damages to crops or facilities) may happen (Dulac-Arnold, Mankowitz, and Hester 2019). Therefore, detailed requirements based on domain knowledge may be defined in advance.

3.3 Slow Growth Rates

As described in Sec. 2, RL is to discover the optimal policy π^* maximizing the expected cumulative reward, so the optimization process accompanies the sequential selection of a new policy π to evaluate. Compared to other RL applications, collecting data in agricultural settings can be slow and expensive, resulting in *data sparsity*. To be specific, observing a transition from s_t to $s_{t'}$ (e.g., recovery from a nutrient deficiency) after an action a_t (e.g., using fertilizers) may take several days or weeks, where $t' > t$. In addition, this lengthy temporal scale can cause the reward signals to be *sparse* in learning since $r_{t\leq\tau<t'}$ would be likely to be zeros. Moreover, the association between the action a_t and the state $s_{t'}$ might not be precisely identified, due to the interference of the actions at times $t < \tau' < t'$. Such an issue has been discussed as a severe problem in RL for long-horizon tasks (Pathak et al. 2017), and we consider that farming belongs to that task family.

3.4 Multimodal Datasets

Precision agriculture could be achieved if an RL algorithm can perform accurate and rapid state estimation using data streams in different formats gathered from *multiple sources*. Modern agricultural systems are equipped with a wide variety of sensors. Specifically, some may be of numeric values or 2D/3D images from soil-/plant-monitoring sensors while others might come from human experts in the form of knowledge graph representing symbiotic dynamics. Naïve methods for unimodal data could not utilize all the available information nor comprehensively assess the agricultural system, and as a result, RL would not succeed, due to incorrectly inferred states.

3.5 Lack of Explainability

The policy in RL is typically parameterized by a complex machine-learning model such as a deep neural network, which is namely regarded as a *black box* that is computationally powerful but lacks the human-understandable interpretation between inputs and outputs (Angelov and Soares 2020). Consequently, when the behaviors of trained controller are subpotimal, the investigation of its erroneous decisions can be difficult. In addition, human farmers could

Table 1: RL for farming: Challenges and Opportunities

Challenges	Opportunities
System Complexity	BO, Exploration Meta RL, Human
Res.&Safety Constraints	Constrained RL, Human
Slow Growth Rates	Model-based RL&Sim2real Exploration, Meta RL, Human
Multimodal Datasets	Multimodal RL
Lack of Explainability	Explainable RL

not fully trust even a successfully trained controller if it keeps suggesting the actions that conflict with their knowledge with no sufficient explanation.

4 Potential Opportunities

In this section, we present the strategies to address each challenge discussed in Sec. 3 (cf. Table 1). Also, Table 2 summarizes the advantages and limitations of each method. Some of the strategies are also active areas of research, e.g., meta-RL (Sec. 4.5), XRL (Sec. 4.8), etc.

4.1 Bayesian Optimization

We suggest to apply a efficient parameter searching method, e.g., Bayesian Optimization (BO) (Frazier 2018), before RL. Because, 1) the control action space A is too large for RL. BO can help to reduce the searching space. 2) RL cannot make sure to work sample efficiently on all control parameters in different temporal scales, e.g., the CO2 supply rate and the light intensity, etc, which are only needed to be set once before the crop planting. It is one-time optimization that is different from RL which needs to make decisions in each step, e.g. hour-scale temperature.

BO is a sequential design strategy for global optimization of black-box functions that do not assume particular functional forms. It works efficiently with only a few samples, when the dimensionality of control parameters is low.

4.2 Constrained RL

Resource and safety constraints are often embedded in applications of RL to agriculture (cf. Sec. 3.2). As discussed in Sec. 2, we suggest using CMDP to take into account *cumulative* constraints and *instantaneous* ones (Liu, Ding, and Liu 2021). The basic idea to solve cumulative constraints is to reduce the constrained optimization problem to an unconstrained problem (Liu, Ding, and Liu 2020b). More specifically, Lagrangian relaxation (Altman 1999) is the most common approach to address cumulative constraints, which outperforms reward shaping that embeds the constraint within the reward function using fixed multiplier λ. As explained in (Tessler, Mankowitz, and Mannor 2018), reward shaping is hard to adapt in environments with different levels of reward values (e.g., high values or low values). Moreover, it is easy to lead the policy to a local optimum. To satisfy instantaneous constraints, an effective approach is to adjust actions at each step by projecting them onto a space for evaluation of feasibility (Liu and Liu 2021). This can be performed by introducing a projection layer into the end of the policy neural network (Liu, Ding, and Liu 2020a). Overall, the constrained RL could learn a significant practical policy at certain cost of computation resource and sample efficiency when the data collection is restricted with cumulative constraints.

4.3 Model-based RL & Sim2Real

Since learning from continuous interactions with a real-world farming environment can be expensive (cf. Sec. 3.3), an alternative idea could be to learn a model of the environment itself to *simulate* the realistic feedback during the control-policy learning (Janner et al. 2019). In this approach, samples from the environment (e.g., a greenhouse) are not fed to learn the controller but instead to the model of the environment. RL can then be performed for policy learning by running the learned environmental model to rapidly respond to a number of actions that the trained controller executes. Such a method would be useful to alleviate the scarcity of observations and rewards discussed in Sec. 3.3.

A related issue can arise due to the simulation-to-real gap. Even though we have a simulator of an environment, the simulated feedback is an approximation of the real environment, so successful policies on the simulator might not generalize well to real scenarios. To close this performance gap, three following approaches can be suggested: 1) *System identification* to build a reliable simulator based on mathematical, physics-based models for a real agricultural system; 2) *Domain adaptation* to shape the data distributions or representations from a simulator to match those in real scenarios (Chen et al. 2021); and 3) *Domain randomization* to create random variants of simulated environmental properties so that this augmented distribution could include plausible observations from real environments (Lee et al. 2019).

4.4 Exploration Strategies

Effective exploration can significantly improve sample efficiency by gathering more informative trajectory data to escape locally optimal solutions in RL while encouraging the controller to visit *novel* state-action pairs. Recently, successful strategies have been proposed under the following categories (Yang et al. 2021): 1) Uncertainty-based methods (Liu and Liu 2023b) estimate the uncertainty (variance) of objective (Eq. 1) via Bayesian posterior to consider it in choosing the next actions; and 2) Intrinsic-motivation methods (Liu and Liu 2023a) take the prediction error of next state as an intrinsic reward to drive the controller towards unfamiliar environments. Similarly, RL controllers in agricultural scenarios could take advantage of a well-designed exploration module because, as pointed out in Sec. 3.1, the state-action space may be extremely large and in Sec. 3.3, sampling can be time-consuming, and reward signals are sparse. One potential limitation can be that the *novel* state-actions pairs may violate *safety* requirements, so a combined solution with constrained RL needs to be considered in practice (Efroni, Mannor, and Pirotta 2020).

Table 2: Pros and cons of different methods

Methods	Pros	Cons
BO	Data-efficiency	Low-dimensional and global control parameters
Constrained RL	Resource and safety constrained controls	Data inefficiency with cumulative constraints Computation complex
Model-based RL Sim2real	Data-efficiency, Time-efficiency	Need well-learned model/simulators Computation complex
Exploration	Data-efficiency, Escape local optima	May explore unsafe actions
Meta RL	Generalization, Adapt to new environment	Data inefficiency during training
Learn from Human	Data-efficiency, Model accuracy	Need human efforts, Less generalization
Multi-model RL	Multi-model data, Information complementary	Need specialized approach to align modalities

4.5 Meta RL

Meta RL (Duan et al. 2016) enables to quantify how different a policy performs on a wide range of learning environments (e.g., different crops) and utilize this experience for learning in novel environments (e.g., new types of crops) to be much *faster* than otherwise possible. In other words, we can expect that after only a mini learning session, a meta RL policy can adapt to new environments that have never been encountered during training time. Meta RL, thus, should be considered to 1) acquire a better generalizable policy in highly complex agricultural environments (cf. Sec. 3.1) and 2) speed up the learning while observations are gathered at a sluggish rate (cf. Sec. 3.3).

4.6 Learning from Human Experts

RL can be particularly hard in farming systems, compared to other applications, because of their complexity (Sec. 3.1), slowness (Sec. 3.3), and safety requirements (Sec. 3.2). To address those issues, existing demonstration data from human experts could be utilized to provide a trained controller with high-quality state-action trajectories in terms of performance and safety. For instance, *imitation learning* (IL) could be designed to supervise a controller to follow the exactly equivalent action to the human choice for each encountered state (Yin et al. 2022). IL can, however, easily fail when the learned policy makes a mistake to deviate from familiar trajectories. *Offline RL* (ORL) can be an alternative method, which approximates the value function of possible state-action pairs (Agarwal, Schuurmans, and Norouzi 2020). In particular, *Conservative Q-Learning* (Kumar et al. 2020) can be a good design choice to learn a safe policy from human demonstrations in agricultural scenarios because it avoids overestimating the values of unseen, risky state-action pairs. Moreover, *reward shaping* (Hu et al. 2020) can be considered to use domain knowledge to promote particular action selections with denser reward signals. For example, as the controller selects a desirable decision according to the domain knowledge, the expert could offer some reward even though its positive consequence is still steps away.

4.7 Multimodal RL

Diverse aspects of state in crops can be captured via *a variety of sensors* into different formats—such as visual images and time-series climate or chemical measurements. Some knowledge from human experts could also be a useful form of data to provide an informative assessment of the plant system based on the observed facts and their known relations. To fully take advantage of all those sources, we suggest considering Multimodal RL (Ma et al. 2022), in which the controller learns a unified policy whose states were generated by both visual and auditory data. In particular, they claim that if one of the modalities is uninformative at some time instants possibly due to too much noise, information from others could be used to compensate for it. Also, their experiments show the importance of developing a specialized approach to align the modalities since simply concatenating all modalities leads to suboptimal performance.

4.8 Explainable RL

The goal of Explainable RL (XRL) (Milani et al. 2022) is to elucidate the decision-making process of controller. For example, XRL in farming scenarios could explain the reason for particular actions chosen such as adding more water or increasing temperature in human-understandable manners. Primary XRL methods fall into three categorizations: 1) Feature-importance explanations identify the features that affect a controller's action choice a_t for the input state s_t, which provide an action-level look at the controller's behavior. 2) Learning process and MDP explanations show the past experiences or the components of the MDP that led to the current behavior, which provide useful information about the effects of the training process or the MDP. 3) Policy-level explanations illustrate the long-term behavior of the controller, which are critical for understanding a controller's behavior to evaluate its overall competency.

5 Conclusion

In this paper, we suggest using Constrained Markov Decision Process to obtain a policy for autonomous farming with reinforcement learning. In addition, we have discussed practical, unique challenges to consider for agricultural applications, for each of which potential solutions also have been presented. We argue that CRL is a greatly promising research direction to realize fully automated operation of farm systems, and our exploration over the topic in this short paper would be a useful resource for researchers and practitioners to perform deeper study on it in the future.

6 Acknowledgments

The work was partially supported through grant USDA/NIFA 2020-67021-32855, and by NSF through IIS-1838207, CNS 1901218, OIA-2134901.

References

Afsar, M. M.; Crump, T.; and Far, B. 2021. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys (CSUR)*.

Agarwal, R.; Schuurmans, D.; and Norouzi, M. 2020. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 104–114. PMLR.

Agostinelli, F.; McAleer, S.; Shmakov, A.; and Baldi, P. 2019. Solving the Rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, 1(8): 356–363.

Ajagekar, A.; and You, F. 2022. Deep Reinforcement Learning Based Automatic Control in Semi-Closed Greenhouse Systems. *IFAC-PapersOnLine*, 55(7): 406–411.

Altman, E. 1999. *Constrained Markov decision processes*, volume 7. CRC Press.

Angelov, P.; and Soares, E. 2020. Towards explainable deep neural networks (xDNN). *Neural Networks*, 130: 185–194.

Cao, X.; Yao, Y.; Li, L.; Zhang, W.; An, Z.; Zhang, Z.; Xiao, L.; Guo, S.; Cao, X.; Wu, M.; et al. 2022. igrow: A smart agriculture solution to autonomous greenhouse control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11837–11845.

Chen, X.-H.; Jiang, S.; Xu, F.; Zhang, Z.; and Yu, Y. 2021. Cross-modal Domain Adaptation for Cost-Efficient Visual Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34: 12520–12532.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Dulac-Arnold, G.; Mankowitz, D.; and Hester, T. 2019. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Efroni, Y.; Mannor, S.; and Pirotta, M. 2020. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*.

FAO. 2022. The State of Food Security and Nutrition in the World 2022. *Food and Agriculture Organization of the United Nations (FAO)*.

Frazier, P. I. 2018. A Tutorial on Bayesian Optimization. arXiv:1807.02811.

Gandhi, R. 2022. Deep Reinforcement Learning for Agriculture: Principles and Use Cases. In *Data Science in Agriculture and Natural Resource Management*, 75–94. Springer.

Gautron, R.; Maillard, O.-A.; Preux, P.; Corbeels, M.; and Sabbadin, R. 2022. Reinforcement learning for crop management support: Review, prospects and challenges. *Computers and Electronics in Agriculture*, 200: 107182.

Hu, Y.; Wang, W.; Jia, H.; Wang, Y.; Chen, Y.; Hao, J.; Wu, F.; and Fan, C. 2020. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33: 15931–15941.

Janner, M.; Fu, J.; Zhang, M.; and Levine, S. 2019. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191.

Lee, K.; Lee, K.; Shin, J.; and Lee, H. 2019. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396*.

Liu, Y.; Ding, J.; and Liu, X. 2020a. A constrained reinforcement learning based approach for network slicing. In *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, 1–6. IEEE.

Liu, Y.; Ding, J.; and Liu, X. 2020b. IPO: Interior-point policy optimization under constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4940–4947.

Liu, Y.; Ding, J.; and Liu, X. 2021. Resource allocation method for network slicing using constrained reinforcement learning. In *2021 IFIP Networking Conference (IFIP Networking)*, 1–3. IEEE.

Liu, Y.; Ding, J.; Zhang, Z.-L.; and Liu, X. 2021. CLARA: A Constrained Reinforcement Learning Based Resource Allocation Framework for Network Slicing. In *2021 IEEE International Conference on Big Data (Big Data)*, 1427–1437. IEEE.

Liu, Y.; Halev, A.; and Liu, X. 2021. Policy learning with constraints in model-free reinforcement learning: A survey. In *The 30th International Joint Conference on Artificial Intelligence (IJCAI)*.

Liu, Y.; and Liu, X. 2021. Cts2: Time series smoothing with constrained reinforcement learning. In *Asian Conference on Machine Learning*, 363–378. PMLR.

Liu, Y.; and Liu, X. 2023a. Adventurer: Exploration with BiGAN for Deep Reinforcement Learning. *Applied Intelligence*.

Liu, Y.; and Liu, X. 2023b. Farsighter: Efficient Multi-step Exploration for Deep Reinforcement Learning. In *The 15th International Conference on Agents and Artificial Intelligence (ICAART)*.

Ma, J.; Chen, Y.; Wu, F.; Ji, X.; and Ding, Y. 2022. Multimodal Reinforcement Learning with Effective State Representation Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1684–1686.

Milani, S.; Topin, N.; Veloso, M.; and Fang, F. 2022. A Survey of Explainable Reinforcement Learning. *arXiv preprint arXiv:2202.08434*.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Roopaei, M.; Rad, P.; and Choo, K.-K. R. 2017. Cloud of things in smart agriculture: Intelligent irrigation monitoring by thermal imaging. *IEEE Cloud computing*, 4(1): 10–15.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Yang, T.; Tang, H.; Bai, C.; Liu, J.; Hao, J.; Meng, Z.; and Liu, P. 2021. Exploration in deep reinforcement learning: a comprehensive survey. *arXiv preprint arXiv:2109.06668*.

Yin, Z.-H.; Sun, L.; Ma, H.; Tomizuka, M.; and Li, W.-J. 2022. Cross Domain Robot Imitation with Invariant Representation. In *2022 International Conference on Robotics and Automation (ICRA)*, 455–461. IEEE.