A Differential Game Model on the U.K. Government and Zero-emission Vehicle Manufacturer: method of Deep Reinforcement Learning

Chuqing Xi*

Department Of Economics, University of Warwick, Coventry, The United Kingdom, CV4 7AL

* Corresponding Author Email: Xicelia962@gmail.com

Abstract. In this paper, a model based on the Differential Game is constructed to solve out the optimized funding value to the Zero-emission Vehicles manufacturers' battery research for the U.K. government. Multi-Agent Deep Deterministic Policy Gradient, a Reinforcement Learning algorithm, is used to solve out the Feedback Nash Equilibria. The optimal model outcome is compared with the empirical data from 2016 to 2023 from Statista and solve out the optimal funding value for the future 12 years. Since the focus of the paper is seeking optimal solutions rather than predicting, we will analyse the action gap between the government and the optimal solutions. Through our analysis, we provide policymakers with actionable recommendations for maximising the effectiveness of government funding in driving innovation and the transition to sustainability. Our findings underscore the importance of strategic patience, an innovation ecosystem, and proactive engagement with the manufacturers on achieving optimal outcomes for the ZEV market and society.

Keywords: Dynamic Optimization, Differential Game, Deep Reinforcement Learning, Environmental Economics.

1. Introduction

Global warming caused by carbon emissions has become a pressing concern. It has propelled more and more governments worldwide to re-evaluate their policies on reducing carbon emissions. Central to this effort is the transition towards sustainable transportation, with zero-emission vehicles (ZEVs) emerging as a key player in reducing carbon emissions. The battery, a prominent element of the ZEV, is one of the limiting factors in the progress, due to the complicated and inefficient process of elements mining and refining. Thus, many governments focused on galvanising the efficiency of battery production. The U.K. government offered funding for the ZEV manufacturers on battery technology innovation to boost ZEV production and demand. Thus, exploring an efficient funding strategy is crucial to the government and society.

Efforts to mitigate environmental pollution have prompted a growing interest among scholars in exploring the optimal strategy for reducing carbon emissions and galvanising the ZEV market. According to Watanabe, Wakabayshi, and Miyazawa [1], investment in green-energy field R&D leads to a technology shock and contributes to a dramatic increase in production; these increases led to a dramatic decrease in price. The “virtuous cycle” is formed between R&D, market growth, and price reduction. Based on this mechanism, this paper will further explore the interaction between R&D funding and price. Gu, Zhou, and Ieromonachou [2] use Stackelberg game theory to build a mathematical model under the condition of incomplete information. They found a way for government to allocate subsidies to ZEV manufacturer that satisfy the profit maximization for all four players in the model: government, ZEV manufacturer, retailer, and consumers in the market. Chen et al [3] built a Stackelberg game model to evaluate the efficiency of the collaboration between vehicle manufacturers and retailers and conclude that collaborative approach promotes the long-term development of supply chain. In line with the research presented above, this paper will focus on profit maximization for the government and manufacturer, but with continuous time structure, complete information assumption, and asymmetric game utility functions, by solving the outcome more quickly using a heuristic method.

Recently, Hu and Laurière [4] tease out the development in machine learning methods for Stochastic control and games, they state the new breakthrough machine learning methods are based on the approaches for high-dimensional partial differential equations, backward stochastic differential equations or on model-free reinforcement learning for Markov decision processes. These unlock the probability to solve complex game with high dimensions which is not feasible with traditional numerical methods. Based on the above papers that prove the feasibility of combining game theory models and machine learning, this paper will use areinforcement learning method to solve a differential game model of players learning from each other and the market environment with stochastic control. Other papers like Wang et al. [5] , Xu et al. [6] and Asgharnia et al. [7] focus on the Pursuit-evasion differential game using DDPG in reinforcement learning while this paper will use MADDPG algorithm which allows continuous action space and multiple players with better convergence. The state transition function will be mutual rather than different individual functions in this paper, this setting allows the model to achieve profit maximization quicker.

2. Model

This model aims to solve the Nash equilibrium to achieve the best responses for players to reach utility maximization. The model is based on the framework of a differential game, which is a game with continuous time setting and rational players. A continuous time framework can better represent reality than a stage game with discrete time since the dynamic changing of the environment can let players update their beliefs instantaneously with no time lag. It includes two players: the U.K. government and ZEV manufacturer. As can be seen in Figure 1, in the first stage of the game, the government will observe the market environment (the ZEV demand at time t, denoted as $\dot{D}(\mathrm{t})$) and make an action $S(t)$ to fund the ZEV manufacturer. After that, ZEV manufacturer will make action $P(t)$ according to $S(t)$ and the $\dot{D}(\mathrm{t})$. The state transition function $\dot{D}(\mathrm{t})$ will automatically transit to the next stage $\dot{D}(t+1)$ according to the actions of player 1 and player 2. Then the government will adjust the amount of funding $S(t+1)$ according to the state transition value $\dot{D}(\mathrm{t}+1)$ … This process goes on repeatedly until both players reach Nash equilibrium or the termination conditions. Variables' descriptions can be seen in Table 3.

- *Assumption 1: Players and actions*

All ZEV manufacturers in the market are sum up to an entire firm in the game. Player 1 is the government. Player 2 is the ZEV manufacturer. The action of player 1 is S, which is the funding value in million pounds to manufacturer on their battery innovation while the action of player 2 is P, which represents the average price of one ZEV in thousand pounds.

- *Assumption 2: Rules of the game*

Both players satisfy individual rationality and incentive compatibility, which means they will maximize the reward function and release their decisions honestly to each other.

- *Assumption 3: Existence of Nash Equilibrium*

The reward functions of both players are concave (linear quadratic form), implying the players are risk averse. Under this condition, the Nash equilibrium exists and is unique in multiagent noncooperative game context. This is proved by Luo and Saigal [8] using Kakutani fixed point theorem and theorem of Gale and Nikaido.

- *Assumption 4: Demand function*

The Demand Q is negatively related to the price and positively related to the funding from the government. It assumes government's effort on battery innovation can be observed by people so that they will believe ZEV market is potentially promising, and thus the demand will increase if the government fund the manufacturer more.

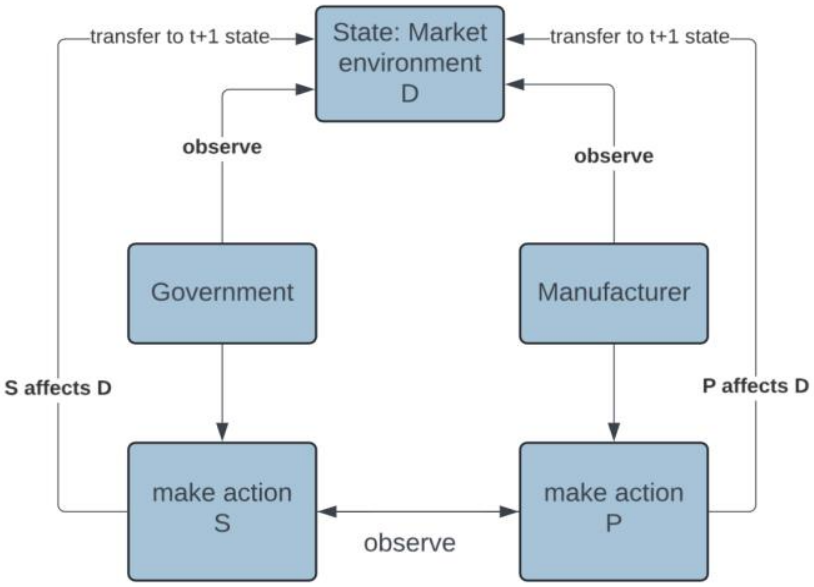

Figure 1. Model mechanism

The market demand $Q(t)$ in period t will be affected by the P and S in the previous period $t-1$ since effect of the demand adjustment on price is not instantaneous and will have a time lag and it will also take time for people to realize the government's efforts on ZEV market. The demand function can be expressed as follows:

$$Q(t) = \alpha - \beta P(t-1) + \psi(t)S(t-1) \tag{1}$$

where $Q(t), P(t-1)$, $\psi(t)$ and $S(t-1)$ is function of t.

- *Assumption 5: Market Environment setting*

Market environment can be described as demand function with a stochastic term to represent the unforeseen exogenous shock. It can be observed by both players and will affect their actions. The stochastic term B follows a Gaussian Distribution, where σ_b is the fluctuation parameter. It can be described as follows:

$$\dot{D(t)} = \frac{dD(t)}{dt} = \frac{[\alpha - \beta P(t-1)]}{dt} + \frac{\psi(t)S(t-1)}{dt} + \frac{\sigma_b dB}{dt} \tag{2}$$

- *Assumption 6: Reward functions*

Government aims to maximize the social reward, it has a reward function contains consumer and producer surplus, it can be described as follows:

$$\mathrm{R_g(t)} = -\gamma \mathrm{S(t)}^2 + \theta \mathrm{S(t)} + \pi_c + \pi_p \tag{3}$$

Where the π_c, π_p are consumer surplus and producer surplus respectively, the expanded form can be found in Table.3.

ZEV manufacturer aims to maximize profit. The reward function contains profit and a portion of the funding it receives from the government since this funding can potentially improve production efficiency, so the manufacturer can benefit. $\sigma^t P(t)$ is the cost of producing one ZEV, and the σ is a term between 0 and. It indicates that as the time goes up, the cost goes down due to the improvement of the battery production technology. The reward function is as follows:

$$\mathrm{R_z(t)} = \mathrm{D(t)} \times (\mathrm{P(t)} - \sigma^t P(t)) + \phi \mathrm{S(t)} \tag{4}$$

The reward functions are set as linear-quadratic form with respect to players' own actions, as Assumption 1 stated.

- *Assumption 7: Target function*

To take the patience into account, both players have a discount rate r. Therefore, their targets are:

$$\max_{s} E[\int_0^T e^{-r_g t} R_g\big(D(t), P(t), S(t)\big)dt] \quad (5)$$

$$\max_{P} \ E[\int_0^T e^{-r_z t} R_z\big(D(t), S(t), P(t)\big)dt] \quad (6)$$

for government and manufacturer respectively, where e^{-rt} is the discount factor. They both aim to maximize their cumulative reward from $t = 0$ to $t = T$.

Table 1. The Model

Players	Player 1: the government	Player 2: the ZEV manufacturer
Action	S(t)	P(t)
Reward function	$\mathrm{R_g(t)} = -\gamma \mathrm{S(t)}^2 + \theta \mathrm{S(t)} + \pi_c + \pi_p$	$\mathrm{R_z(t)} = \mathrm{D(t)} \times (\mathrm{P(t)} - \sigma^t P(t)) + \phi \mathrm{S(t)}$
Target function	$\max_{s} \mathrm{E}[\int_0^T \mathrm{e}^{-r_g t} \mathrm{R_g}\big(\mathrm{D(t)}, \mathrm{P(t)}, \mathrm{S(t)}\big)\mathrm{dt}]$	$\max_{P} \ \mathrm{E}[\int_0^T \mathrm{e}^{-r_z t} \mathrm{R_z}\big(\mathrm{D(t)}, \mathrm{S(t)}, \mathrm{P(t)}\big)\mathrm{dt}]$
State Transition function	$\dot{D(t)} = \frac{dD(t)}{dt} = \frac{[\alpha - \beta P(t-1)]}{dt} + \frac{\psi(t) S(t-1)}{dt} + \frac{\sigma_b dB}{dt}$	

3. Equilibrium analysis

There are two common types of Nash equilibria in differential game. One is Open-loop Nash equilibrium, it is the one for which each player chooses his control variable for each point in time t at the outset of the game, which is a planned time path of the actions. However, this does not give much implication since it is pre-planned and not realistic. Thus, this paper will focus on the other type, which is called Feedback Nash equilibrium.

3.1. Feedback Nash Equilibrium

Feedback Nash equilibrium allows players to best response in each point of time. Thus, it has property of subgame perfect. Feedback Nash equilibrium is more popular and realistic but with more complicated calculation. To reach Feedback Nash equilibrium, it should satisfy Assumption 7, subject to the state transition equations:

$$\dot{D(t)} = \frac{dD(t)}{dt} = \frac{[\alpha - \beta P(t-1)]}{dt} + \frac{\psi(t) S(t-1)}{dt} + \frac{\sigma_b dB}{dt} \quad (7)$$

And we said their actions are in Nash equilibrium if each player is doing the best given the other's strategy.

3.2. Methodology

Multi-agent deep deterministic policy gradient (MADDPG), a deep reinforcement learning (DRL) algorithm designed for multi-agent scenarios based on DDPG [9], is used for solving Feedback Nash equilibrium.

- *Markov Game: general structure*

A Markov game for 2 agents is defined by a set of states $\boldsymbol{S}$ describing the possible configuration of the two agents, a set of actions $\boldsymbol{a_1}, \boldsymbol{a_2}$ and a set of observations $\boldsymbol{o_1}, \boldsymbol{o_2}$ for each agent [9]. To choose an action, each agent uses the policy π (4.2.2) which produces the next state according to the state transition function in Table 1. Each agent obtains rewards as a function of the state and agent's actions and receives a private observation $\boldsymbol{o}$ correlated with the state. Each agent aims to maximize its own total expected return, which corresponds to the target function in Table 1.

- *Deep Deterministic Policy Gradient (DDPG): how to choose players' actions*

Let us denote by $\pi(s)$ a differentiable deterministic policy. Since a greedy policy becomes problematic in continuous action space [10], DDPG moves the policy in the direction of the gradient of Q where Q is the action-value function, which corresponds to the reward function in Table 1 and y^j in Algorithm 1 below. π takes a state and returns a probability distribution over actions (Zai,

2020, p. 91) [11], where a higher probability indicates an action that is more likely to result in the highest reward.

- *Actor-Critic Methods: how actions and states interact*

The actor refers to the policy, and the critic refers to the estimation of value function Q [10]. In DRL, both the actor and the critic can be represented by non-linear neural network function approximators. As can be seen in Figure 2, state in period t is input to an actor network (policy function π in 4.2.2) and it produces an action in t and a state value for $t+1$. Action exploration depends on randomness. We sample from a distribution to get an action, such that the most probable action is most likely to be sampled, and this allows the transition of the state. Then, the rewards and the state in $t+1$ is generated and input to the Critic network to compute the advantage value (relative value of state $t+1$) and the value of Q function, which are used to reinforce the action the Actor takes.

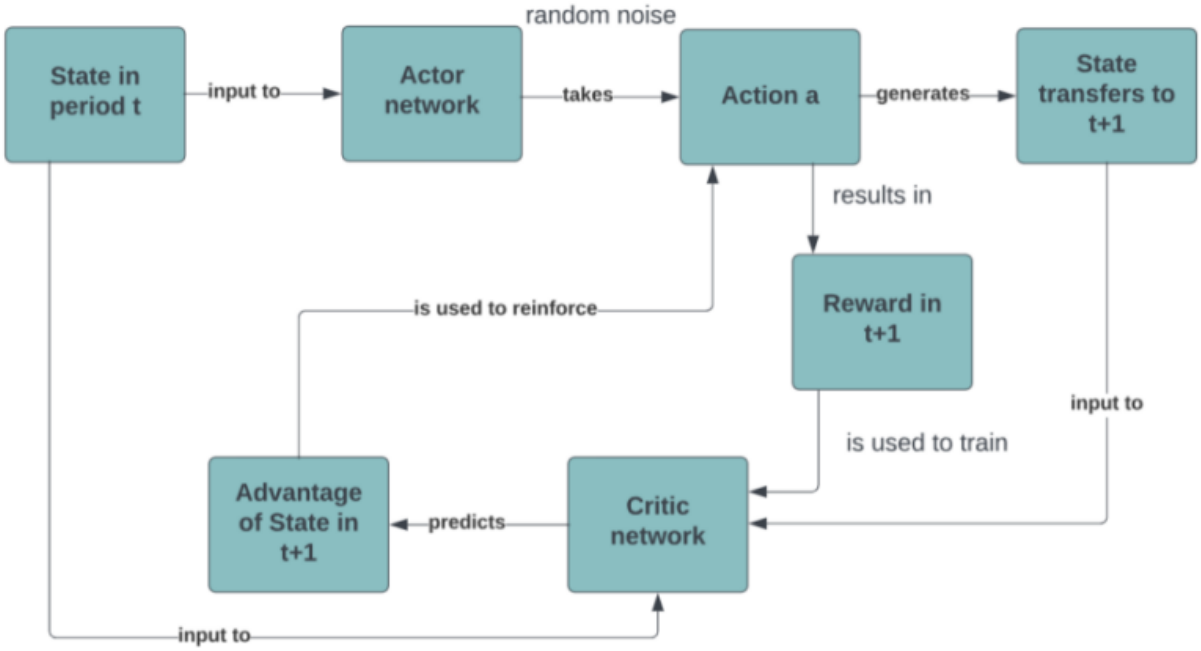

Figure 2. Process of the Actor-Critic Model

- *MADDPG: training process*

As can be seen in Algorithm 1, player 1 and player 2 are treated as two agents with policies parameterized by $\theta = \{\theta_1, \theta_2\}$, and let $\mu = \{\mu_1, \mu_2\}$ be the set of all agent policies. After the actor π predicts the best action and chooses the action to take, which generates a new state, the critic network computes the value of the old state and the new state. At the end of the episode, we compute the return of the episode, which is essentially the sum of the discounted rewards in this episode. Every episode will have a different path of action exploration; they will be stored in the replay buffer for iteration so that the agent can retrieve the data from it and learn from the experience. This random sampling helps reduce the correlation among sequential experiences and stabilizes training.

Algorithm 1: Multi-Agent Deep Deterministic Policy Gradient [9]

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer $\boldsymbol{R}$
for episode = 1, M **do**
Initialize a random process $\mathcal{N}$ for action exploration
Receive initial observation state x
for t = 1, T **do**
Select action $a_i = \mu_{\theta_i}(o_i) + \mathcal{N}_t$ according to the current policy and exploration noise
Execute action $a = (a_1, \ldots, a_N)$ and observe reward r and observe new state x'
Store transition (x, a, r, x') in R

$$x \leftarrow x'$$

For agent $i = 1$ to N **do**
Sample a random minibatch of S samples $\left(x^j, a^j, r^j, x'^j\right)$ from R
Set $y^j = r_i^j + \gamma Q'(x', a_1', \ldots a_N')|_{a_k'=\mu_k'\left(o_k^j\right)}$
Update critic by minimizing the loss: $L = \frac{1}{S}\sum_j \left(y^j - Q(s_i, a_i|\theta^Q)\right)^2$
Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta\mu}\mu(s|\theta^\mu)|_{s_i}$$

end for
Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}$$

end for
end for

Where **o** is the observation value (state value), **a** is the action, and **s** is the state. τ is a parameter called temperature that scales the probability distribution of actions in π. A high τ will cause the probabilities to be very similar, whereas a low τ will exaggerate differences in probabilities between actions [11].

3.3. Implementation

According to the U.K. government [12], the highest registration number of petrol cars in the past 5 years is around 1500,000. Thus, we assume that when the demand for ZEV reaches the same level as that for petrol cars by 2035 and has an increasing trend or is stable around 1500,000, the ZEV mandate is satisfied. We define agents, actions, reward functions, and the state transition function according to Table 1 for DRL environment settings. Our policies are parameterized by a three-layer rectified linear unit multilayer perceptron with 32 units per layer for the actor and critic. Orthogonal initialization was applied to the weights of neural network layers and set biases to zero. It helps in maintaining the variance of the inputs across layers, which can be beneficial for deep neural networks by preventing vanishing or exploding gradients. Set the replay buffer size to 1,000,000 for storing the iterations. Set the hyperparameters and the parameters of the reward functions as below:

Table 2. Hyperparameter and parameter settings

Hyperparameter	Value	Parameter	Value
Learning rate of actor	0.01	α	7800
Learning rate of critic	0.01	β	169
Batch size	1024	ψ	$0.5t$
Evaluate frequency	3	γ	0.4
τ	0.01	θ	20
Buffer size	1000000	ϕ	0.2
Noise exploration	0.8	σ	0.85
Episodes	260	σ_b	2

The adjustments to other hyperparameter settings are based on Lowe et al.[9]. The values of α and β are based on the adjustment of the regression of demand and price per ZEV, using data in table 3. Since the relation between S, P and D is unclear and may change dynamically over time, we tried different values of the parameters in reward functions and trained our models until the action of player 2 (ZEV price) was in line with the past data (see Table 4) generally with the condition of rewards convergence, and then evaluated them by averaging various metrics for around 100 further iterations. The terminal condition is when time step reaches 18, implying year of 2035.

Table 3. Table of the variables and parameters

Variable	Description	Variable	Description
S	Funding from government to ZEV manufacturer	P	Average price of ZEV that the manufacturer sets
π_c	Consumer surplus, equals to $D * (42 - P)$ Where 42 is the maximum price	π_p	Producer surplus, equals to $D * (P - \sigma^t C_{pv})$. Simply the demand times profit per vehicle
σ	Discount factor of the cost per ZEV.	σ_b	Fluctuation parameter of the Brownian motion
D	Demand of the ZEV	B	Brownian motion with 0 mean
r_g	Discount rate, between 0 and 1 that controls how much government discounts future rewards when making a decision.	r_z	Discount rate, between 0 and 1 that controls how much ZEV manufacturer discounts future rewards when making a decision.
R_g	Reward function of the government	R_z	Reward function of the ZEV manufacturer

Table 4. Data (Statista,2023) [13] and (GOV.UK, 2023) [12]

Year	Average EV price in thousand GBP	UK EV Unit Sales (Demand) in thousand vehicles	UK government funding in million GBP
2016	41.554	38.12	\
2017	41.396	48.17	\
2018	41.554	59.96	40
2019	40.685	74.4	25
2020	40.132	174.74	23
2021	40.527	311.5	10
2022	39.895	368.9	25
2023	39.895	376.8	10

4. Result

4.1. Optimal Outcome

Table 5. Action settings

RL/Differential game	Player	Action
Agent 0 (Blue line)	The UK government	Funding value in million pounds
Agent 1 (Orange line)	The ZEV manufacturer	Price per ZEV in thousand pounds

Let us set two different discount rates settings: Model 1 has $r_g = 0.95, r_z = 0.95$ while Model 2 has $r_g = 0.99, r_z = 0.95$ for comparison. As can be seen in Table 5, the blue line represents the best response for the government, while the orange line represents the best response for the ZEV manufacturer. For Model 1 in Figure 3 and 5, the discount rates of both agents are set to 0.95, which suggests that future actions and payoffs are valued at 95% of their immediate value per time step into the future. Conversely, Model 2 assigns a discount rate of 0.99 to the government, implying the future payoffs are nearly equivalent to their immediate values, while maintaining the ZEV manufacturer's rate at 0.95. The higher the discount rate, the less myopic the player is, reflecting a strategic preference for long-term outcomes.

When we compare Figure 3 and 4, we can observe that the government's funding values fluctuate over time, potentially attributable to the stochastic policy action exploration. Nevertheless, optimal funding allocations tend to converge around 39.6 million pounds in Figure 3 and 39.8 million pounds in Figure 4. It suggests that under Model 2, the government amplifies its investment to bolster the perceived value of long-term gains.

Figure 3. Actions when $r_g = 0.95, r_z = 0.9$

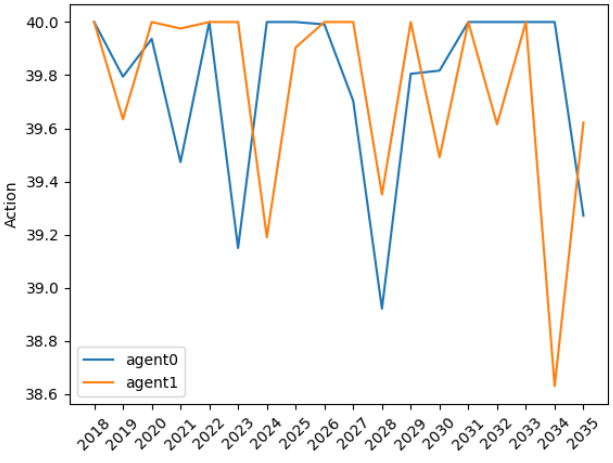

Figure 4. Actions when $r_g = 0.99, r_z = 0.95$

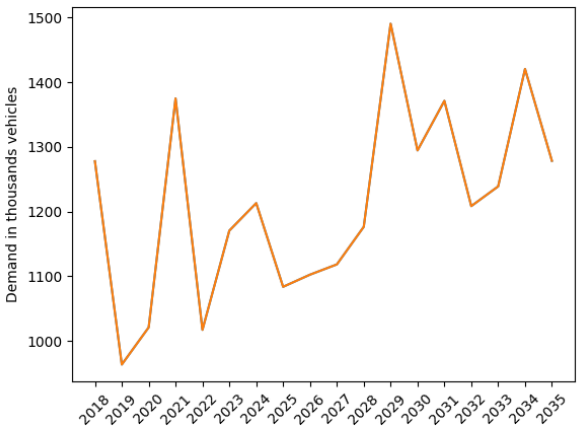

Figure 5. ZEV demand when $r_g = 0.95, r_z = 0.95$

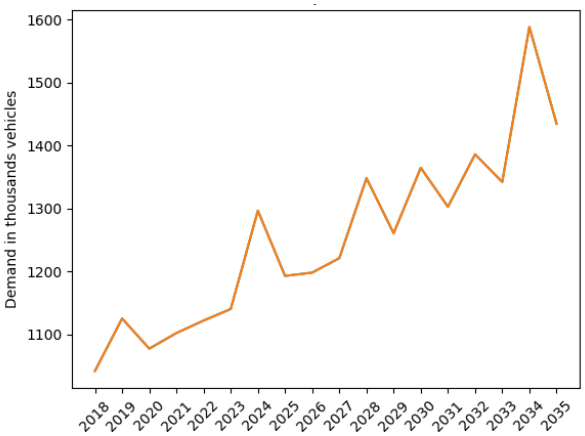

Figure 6. ZEV demand when $r_g = 0.99, r_z = 0.95$

For the manufacturer, the best price settings are centred around 39.1 thousand pounds and 39.7 thousand pounds respectively. Lowering price may catalyse an immediate surge in demand in the short run, as depicted in the initial stages in Figure 5, where the quantity demand is higher than that of Figure 6. However, this trend reverses in the long run, with Model 2 demonstrating superior quantity demand.

As shown in Figure 6, the peak of the demand is approximately 1.6 million in 2034, which satisfies the ZEV mandate set by the government. However, in Figure 5, it indicates for Model 1, the ZEV mandate is unattainable since the quantity demand for ZEV is lower than the threshold of 1.5 million. Compare Figure 5 with 6, it can be observed a more consistent upward trend in ZEV demand within Model 2.

4.2. Model Evaluation

As Lowe et al. mentioned in 2017 [9], we can evaluate the model by checking the convergence of the rewards and the stability of cumulative rewards over episodes. As can be seen in the following figures:

The numerical solutions of the Feedback Nash equilibria demonstrate that the action values can bring both players to a steady state where no one can benefit from deviating to any other actions. This is also called the Pareto outcome. When we observe convergence in Figures 7 and 9, it implies the steady state is reached.

In Figure 7, the rewards for both players converge and reach the steady state at around 2026. In Figure 8, the government's reward converges more slowly than the manufacturer's at around 2035. Convergence of the reward indicates models successfully solves the best actions for both players. The Feedback Nash Equilibria exist and are around 39.6 million (action value) in Model 1 and 39.8 million in Model 2 for the government, while for the manufacturer it is 39.1 thousand pounds and 39.7 thousand pounds respectively. A higher discount rate for the government in Model 2 correlates with a higher cumulative reward over time, suggesting that a more patient government could potentially get a higher payoff.

Figures 8 and 10 demonstrate the effectiveness of the training over episodes, as evidenced by the reward stabilization. In the initial episodes, the reward is high for the government but low for the manufacturer. This is because the neural network explores the action from value 0 at beginning, which implies the government pays nothing and gets a higher reward when the manufacturer sets price to 0 and has no profit. Then, at around the 60th episode, models began to converge to a policy that consistently yields probable highest reward where both players play positive actions. For Model 1, the cumulative rewards for both players enter in stable stage at the 60th episode while for Model 2, the cumulative rewards are further better off at around the 160th episode, achieving a higher reward, indicating a better training outcome.

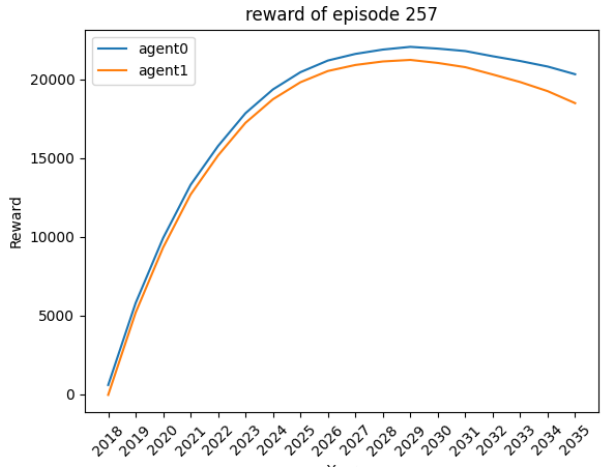

Figure 7. Reward across years

$r_g = 0.95, r_z = 0.95$

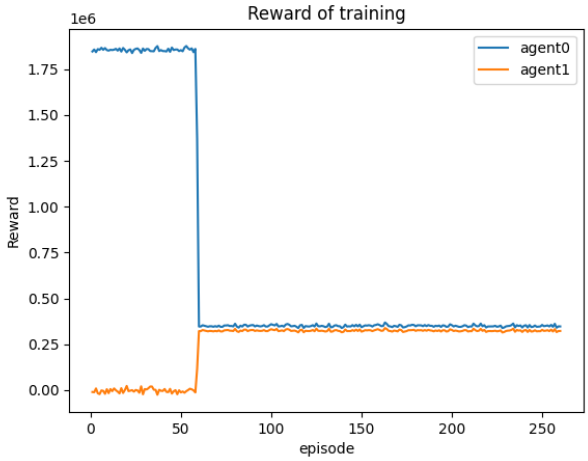

Figure 8. Cumulative rewards for 18 years

$r_g = 0.95, r_z = 0.95$

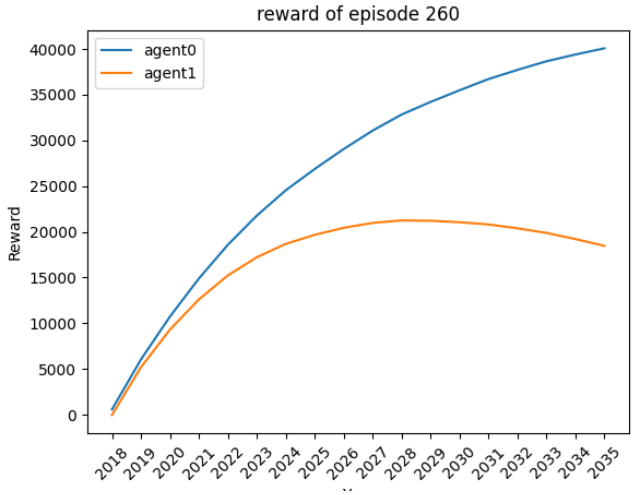

Figure 9. Reward across years $r_g = 0.99, r_z = 0.95$

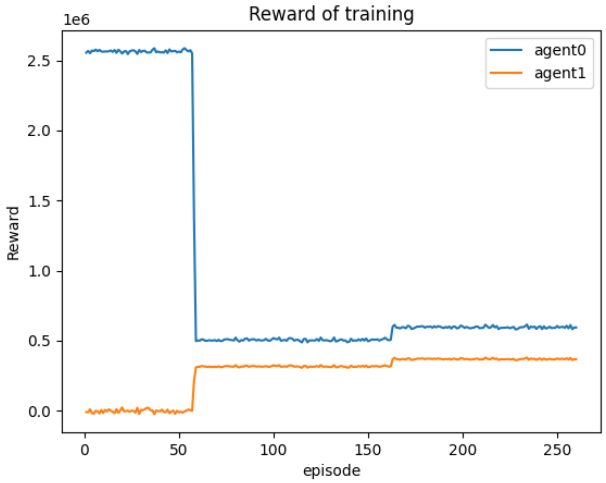

Figure 10. Cumulative rewards for 18 years $r_g = 0.99, r_z = 0.95$

5. Discussion

Although the models focus on solving the optimal outcome rather than predicting real-world behavior, it is still useful to compare reality with the model outcomes. In terms of funding value, the government's funding to ZEV manufacturer was 40 million pounds in 2018, which is matched with Figures 3 and 4. However, it decreased to 10 million pounds in 2021 and increased to around 25 million pounds in 2022, which is less than the optimal action. For the action of player 2, the optimal values are lower than the past data (Table 4) but diverge within a small range. For quantity demand, the optimal value is not accurate compared with the past data. It may be due to the complexity of the ZEV demand function in the real life, which can not be captured by this model.

We observe there are gaps between the model optimal outcome and the past empirical data, meaning the government did not act rationally. One possible explanation is that policymakers are more risk-averse with less patience and are 'boundedly rational', meaning that some constraints apply to knowledge of future events and impacts. Many of the limits on rational policymaking stem from a lack of knowledge, which prevents the evaluation of possibly better alternatives. In many circumstances, it cannot be solved by improved policy analysis since it is imposed by the fact that it refers to future occurrences and actions that are difficult to model or anticipate (Nair and Howlett, 2017) [14]. In the ZEV context, battery innovation is unpredictable, and it will take time to observe and analyze the market. Thus, the government reduced or delayed the optimal funding to evaluate and wait for the innovative process before making subsequent funding decisions. Consequently, it could cause policy failure due to inaction, inaction postponed, or ineffective policy action because of incomplete knowledge and unpredictability. Another possible reason for reducing funding is that not many manufacturers meet the government's criteria. Thus, we suggest that UK policymakers enhance collaboration with manufacturers. Regular communication and collaboration with manufacturers can help manufacturers understand their perspectives. The government can develop an innovation ecosystem to provide incentives for research collaboration, foster entrepreneurship, or adjust the funding criteria to incentivize and galvanize more manufacturers to participate in the challenge. The manufacturers can learn from each other during the competition for funding to some extent, which may improve the product quality. Additionally, a flexible policy scheme could also help reduce the risk. Despite the inherent challenges of predicting future occurrences, the government can gain more accurate insights by enhancing data collection and market analysis. The uncertainty can be alleviated by setting different policies for different potential situations in the future.

6. Conclusion

We built two differential game models for the government and ZEV manufacturer and solve the Feedback Nash Equilibrium by MADDPG. Model 2 performs well with converged rewards and higher cumulative rewards for both players. The feedback Nash equilibrium in steady state is where the government funds 39.8 million pounds to the ZEV manufacturer and the manufacturer sets the price at 39.7 thousand pounds. Under these conditions, ZEV demand is expected to reach around 1600 thousand by 2034. However, we found the government does not react rationally as the model predicted because they are more risk averse. When facing with the uncertainty of technological innovation, they may delay the optimal action to wait and gain more information first. To better achieve their mandate, we suggest the government incentivize more manufacturers to participate in the challenge, adjust their funding criteria for manufacturers, develop an innovation ecosystem, and improve their analysis of the ZEV market.

The first limitation of this paper is that the assumptions of the model may not match the reality. Secondly, since MADDPG depends on neutral networks, it can only solve the approximate numerical solutions rather than the analytical solutions of the Feedback Nash Equilibria, which may not be strictly precise in theory. Moreover, a common issue in DRL training is that since the data is simulated and collected by the agent itself, the results of each collection will change due to changes in seed and hyperparameters. Coupled with the randomness of stochastic policy, this results in the instability of the data. One method to solve this is to pile up the training episodes, but due to the limited computing resources, the maximum episode is 260 in this paper. Further study is suggested to enhance the DRL training and to include more players who could potentially affect the interrelationship in the ZEV market.

Reference

[1] Watanabe, C., Wakabayashi, K., Miyazawa, T. Industrial dynamism and the creation of a "virtuous cycle" between R&D, market growth and price reduction: The case of photovoltaic power generation (PV) development in Japan[J]. Technovation, 2000, 20(6): 299-312.

[2] Gu, X., Zhou, L., Ieromonachou, P. Subsidising an electric vehicle supply chain with imperfect information[J]. International Journal of Production Economics, 2019, 211: 82-97.

[3] Cheng, Y., Yao, Z., Wang, X. Differential Game Analysis for Cooperation Models in Automotive Supply Chain Under Low-Carbon Emission Reduction Policies[C]. In: International Workshop on Frontiers in Algorithmics, Cham: Springer Nature Switzerland, 2023, 147-159.

[4] Hu, R., Laurière, M. Recent developments in machine learning methods for stochastic control and games [DB/OL]. arXiv preprint, 2023, arXiv:2303.10257.

[5] Wang, M., Wang, L., Yue, T. An Application of Continuous Deep Reinforcement Learning Approach to Pursuit-Evasion Differential Game[C]. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019: 1150-1156.

[6] Xu, C., Zhang, Y., Wang, W., Dong, L. Pursuit and evasion strategy of a differential game based on deep reinforcement learning[J]. Frontiers in Bioengineering and Biotechnology, 2022, 10: 827408.

[7] Asgharnia, A., Schwartz, H., Atia, M. Learning deception using fuzzy multi-level reinforcement learning in a multi-defender one-invader differential game[J]. International Journal of Fuzzy Systems, 2022, 24(7): 3015-3038.

[8] Luo, Q., Saigal, R. Dynamic multiagent incentive contracts: Existence, uniqueness, and implementation [J]. Mathematics, 2020.

[9] Lowe, R., Wu, Y.I., Tamar, A., Harb, J., Pieter Abbeel, O., Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments[C]. Advances in Neural Information Processing Systems, 2017, 30.

[10] François-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., Pineau, J. An introduction to deep reinforcement learning[M]. Foundations and Trends in Machine Learning, 2018, 11(3-4): 219-354.

[11] Zai, A., Brown, B. Deep reinforcement learning in action[M]. Manning Publications Co., 2020: P10-1.

[12] GOV.UK. Vehicle licensing statistics: 2022[DB/OL]. Available at: https://www.gov.uk/government/statistics/vehicle-licensing-statistics-2022/vehicle-licensing-statistics-2022 [Accessed 2023].

[13] Statista. Electric Vehicles – UK. Statista Market Forecast [DB/OL]. Available at: https://www.statista.com/outlook/mmo/electric-vehicles/united-kingdom?currency=usd#unit-sales [Accessed 2023].

[14] Nair, S., Howlett, M. Policy myopia as a source of policy failure: Adaptation and policy learning under deep uncertainty[J]. Policy & Politics, 2017, 45(1): 103-118.