The Research on Evasion Strategy of Unpowered Aircraft Based on Deep Reinforcement Learning

Keqin Chen¹, Jianchang Lei², Bin Li³*

¹China Academy of Launch Vehicle Technology, Beijing, China
²China Academy of Launch Vehicle Technology, Beijing, China
³China Academy of Launch Vehicle Technology, Beijing, China
* Corresponding author: 04nwpu@163.com

Abstract—For the pursuit-evasion (PE) game, this paper proposes an evasion strategy with coordinated control of angle of attack, bank angle and body morphing control using deep reinforcement learning. Considering the evasion and ballistic regression of the aircraft, the specified miss distance (SMD) and residual energy are used as the optimization objectives, to acquire the optimal control strategy against the encounter with pursuer in the terminal guidance phase. For the problem of sparse rewards, reward reshaping cannot be performed for this problem, we modify DQN algorithm with the mechanism of Monte-Carlo reinforcement learning to improve the sampling efficiency and realize the end-to-end learning. Finally, the linear analytical solution of the problem based on SMD is analyzed theoretically. With it, the strategy obtained by reinforcement learning is compared and explained.

1. INTRODUCTION

For the PE game problems, the usual method is to linearize the confrontation model, use the method of optimal control [1-2] or the differential game [3] to solve it. Through these algorithms, a certain theoretical solution can be derived. However, for the optimal control method, it is necessary to use SQP [4], the pseudospectral method [5-6], or the target shooting method [7-8] to solve problems offline under the consideration of nonlinear models. The method of the differential game does not need the real-time state of both parties but the overload limit value of both sides. This strategy is often too conservative and usually leads to redundant maneuver, which has impacts on continuing to complete other tasks after evasion. Considering the ballistic regression and target guidance after evasion, Weiss et al. [9] proposed a more practical solution to the pursuit strategy and derived the most energy-saving pursuit strategy based on the specified miss distance (SMD).

With the artificial intelligence achievements in Go with Deep Mind [10], deep reinforcement learning has been paid more and more attention. Some scholars have applied deep reinforcement learning to the generation of evasion strategies [11-12]. However, these models almost adopt one similar idea, that is, the method of Macro Actions [13], which is based on the various forms of maneuvering and certain experience summarization of aircraft. However, for high-speed unpowered aircraft, when the sensor limits the detection distance, the execution time of the evasion strategy is only tens of seconds, and it is difficult to make complex maneuvers. Attack of angle (AOA) and bank angle must be controlled accurately [14], to implement end-to-end reinforcement learning. Gaudet et al.
studied the application of meta-reinforcement learning for high-speed aircraft under multiple problems and achieved some good results [15-18].

For this PE problem, the following distinct characteristics exist:
1. Compared with the PE game outside the atmosphere [19], there is no need to consider the constraints of carrying fuel, only the constraints of heat flow, overload, etc. [20];
2. It is difficult to give a certain prior knowledge strategy during the round. Only at the end of the round can the miss distance be obtained to measure the results;
3. Due to the high relative speed of the high-speed aircraft, the lateral maneuvering ability of both sides is greatly restricted, and the encounter time can be approximately predicted [21].

Given the above characteristics, only sparse rewards can be given, which reward cannot be given before the miss distance is obtained. For the common sparse reward problem, there are two kinds of solutions. One is the transformation of the reward function, which applies professional knowledge to set dense rewards. But the reward function may be misleading. It is easy to fall into the local optimum. The other is to apply modifications to the algorithm. Curiosity [22], PER [23], HER [24], and other algorithms are all effective solutions to sparse problems and do well in the search for rewards. However, the reward can be obtained after a predictable time. There is no need to consider how to search for rewards, only the efficiency of sampling needs to be improved. So, we borrow the mechanism of Monte-Carlo reinforcement learning to modify the DQN algorithm for making every training valuable.

This paper establishes a nonlinear confrontation model between a high-speed aircraft with morphing control ability and a pursuer in the terminal guidance. We adopt the modified DQN algorithm to minimize the energy loss, setting the miss distance as a constraint. Considering the control variable constraints, a network with a certain evasion strategy is trained, whose parameters can be configured on the aircraft guidance system to realize online decision-making. Finally, the control strategy is compared with the linearized model to illustrate the similarity between the linear model and the nonlinear model.

2. Establishment of the PE Game Model

2.1 Kinetic Equations
Assuming that the high-speed unpowered aircraft is a mass point, the influence of the earth's rotation is not considered, and the sideslip angle is set to 0. The three-degree-of-freedom (3DOF) dynamic model of the unpowered aircraft is:

\[
\begin{align*}
\dot{v} &= -\frac{D}{m} - g \sin \theta \\
\dot{\theta} &= \frac{1}{v} \left( L \cos \varphi - g \cos \theta \right) \\
\dot{\psi} &= \frac{L \sin \varphi}{m v \cos \theta}
\end{align*}
\]

In the formula: \( \theta \) is the aircraft speed inclination angle, \( \psi \) is the heading angle of the aircraft, \( D \) and \( L \) are the drag and lift of the aircraft, respectively, \( v \) is the flight speed, \( \varphi \) is the bank angle of the aircraft, and the mass of the high-speed aircraft is 907kg, the reference area is taken as \( 0.35 \ m^2 \).

The 3DOF dynamic model of the pursuer is:

\[
\begin{align*}
\dot{v} &= \frac{T \cos \alpha - D}{m} - g \sin \theta \\
\dot{\theta} &= \frac{1}{v} \left( T \sin \alpha \cos \varphi + L \cos \varphi - g \cos \theta \right) \\
\dot{\psi} &= \frac{T \sin \alpha \sin \varphi + L \sin \varphi}{m v \cos \theta}
\end{align*}
\]

2
In the formula, $\alpha$ is the AOA of the aircraft, $P$ is the thrust of the engine, which is generally in the terminal guidance phase $P = 0$. The kinematic equations are the same for the escaper and the pursuer:

\[
\begin{align*}
\dot{x} &= v \cos \theta \cos \psi \\
\dot{y} &= v \sin \theta \\
\dot{z} &= -v \cos \theta \sin \psi \\
\end{align*}
\] (3)

In the formula: $x$, $y$, and $z$ are the coordinates of the aircraft in the ground coordinate system. In the atmosphere, the general pursuer adopts pure proportional navigation (PPN) to generate lateral command acceleration for pursuing the target.

\[
\begin{align*}
n_{x} &= Nv_d \dot{q}_d \cos \Delta \psi \\
n_{z} &= Nv_d \dot{q}_d \cos \theta \sin \psi + Nv_m \dot{q}_m \sin \Delta \psi \sin \theta \\
\end{align*}
\] (4) and (5) are respectively the forms of PPN under the ballistic system. In the formula: $n_x$ and $n_z$ are the command acceleration, $\dot{q}_d$ and $\dot{q}_m$ are the line-of-sight (LOS) pitch angle and yaw angle rate of change, respectively, $N$ is the navigation ratio and $\Delta \psi$ is the angle between the speed of the pursuer and the relative speed at the projection on the ground.

### 2.2 Analysis of Objectives

Aiming at the maximum miss distance is not suitable for aircraft with extremely strict energy management such as an unpowered vehicle. If the miss distance is greater than the capture radius, the goal of evasion can be achieved. In this case, the value of the miss distance has no effect on the result however large the miss distance is. But after the evasion, the aircraft has to return to its mission. It is necessary to pursue as little energy consumption as possible to ensure that there is enough energy to complete the final mission. The above analysis shows that this problem is a kind of SMD problem [9], to which the optimal control method is applied for solving it.

The optimized goal can be expressed as:

\[
J = \max E(t_f) \\
s.t. r_{\text{miss}} > r_h \\
\] (6)

Where, $E(t_f)$ is the residual energy at the end of the game, $r_{\text{miss}}$ is the miss distance, and $r_h$ is the capture radius of the pursuer.

### 3. METHODOLOGY

For nonlinear problems, it is often impossible to obtain analytical solutions directly by theoretical analysis. To obtain the evasion strategy with adaptive ability, we adopt reinforcement learning as a strategy generation method.

#### 3.1 Monte-Carlo Deep Reinforcement Learning

Reinforcement learning is an optimization algorithm based on the "trial and error" principle. Through the process of exploration-learning-re-exploration, learning of the environment is carried out. For each action $a$, a certain reward $r$ can be given to evaluate the quality of the action. According to the different ways of updating policy, it is divided into policy update and value function update.

The common algorithm for value function update is Q-Learning, which stores a Q table to represent the current strategy, and updates the value in the Q table to achieve the update of the strategy. For the problem of large state space, it is difficult to record the state-action Q value through a Q table. Therefore, the DQN algorithm is proposed to replace the Q table with a deep network that realizes the learning of the relationship between state-action and Q value. To promote the efficiency of training, replay buffer technology is used.
However, both Q-learning and DQN algorithms use the TD update method, that is, one training is performed at each step of the simulation, which leads to the inefficiency of learning in the case of sparse rewards. The Monte-Carlo (MC) method uses an episode-learning-based method to update the strategy. This method updates \( Q(s, a) \) incrementally with (7).

\[
Q(s, a) = \frac{c(s, a)Q(s, a) + R}{c(s, a) + 1}
\]  

(7)

Where, \( c(s, a) \) is the sampling times of \( (s, a) \). (7) can be further simplified to (8).

\[
Q(s, a) = Q(s, a) + \xi(R - Q(s, a))
\]

(8)

Where, \( \xi = 1/(c(s, a) + 1) \). When \( \xi \) is set as hyper-parameter, the policy update speed can be controlled, which can be defined as the learning rate. As for the recording of \( Q(s, a) \) needs Q-tables, it is difficult to solve the problem in the case of large state and action space. We can parameterize \( Q(s, a) \) into \( Q(s, a; \theta) \) with a neural network. From (8), it can be concluded that the objective function of policy update is:

\[
J(\theta) = (R - Q(s, a; \theta))
\]

(9)

The corresponding parameter \( \theta \) update method is:

\[
\theta = \theta + \xi(R - Q(s, a; \theta))\nabla_{\theta}Q(s, a; \theta)
\]

(10)

At the same time, the replay buffer can greatly improve the sample rate and solve the problem of slow updating caused by episode-learning. The algorithm is illustrated in Table 1. For the problem of sparse rewards, the strategy can be updated after the reward is obtained, reducing meaningless policy updating and improving the efficiency of training.

| TABLE 1. | MC-DQN ALGORITHM |
|----------------|------------------|
| **MC-DQN algorithm** |     |
| Initialize the experience pool \( D \) with a capacity of \( N \) |
| Initialize the neural network with random weights \( \theta \) |
| For episode = 1, N_MAX do |
| Initialize the environment |
| For t=1, T do |
| Select actions in state according to \( \varepsilon \)-greedy strategy |
| Perform action \( a_t \) and get reward \( r_t \), the next state \( x_{t+1} \) from the environment |
| Record status pair \( <s_t, a_t, r_t> \) |
| End For |
| Initialize the sequence \( G \) to zero, and set \( G_0 = r_0 \) |
| For t=T-1,1 do |
| \( G_t \leftarrow \gamma G_{t+1} + r_t \) |
| Store data pair \( <s_t, a_t, G_t> \) into the experience pool \( D \) |
| End For |
| Sample small batches of data \( <s_t, a_t, r_t> \) with batch size from the experience pool \( D \) |
| Let \( y_t = G_t \), perform gradient descent algorithm on \( (y_t - Q(s, a; \theta))^2 \) |
| End For |
3.2 Analysis of the Optimal Evasion Strategy

Considering the strategies obtained by reinforcement learning are unexplainable, we make a linear analysis of this problem. Because of the BTT control of the aircraft in three-dimensional space, the pitch channel, yaw channel, and roll channel are quite different. It is difficult to directly find the evasion strategy against the pursuer based on proportional guidance. But the optimal maneuver direction of the escaper is perpendicular to the initial collision surface [25], that is, maneuvering in the horizontal plane. Meanwhile, when the leading angles of both sides are less than 45°, The 2D and 3D results are approximated [25]. The confrontation can be approximately viewed as in the horizontal plane. Consequently, the PE game in the horizontal plane is considered.

For the PE game problem in two-dimension, the state variable is set as

\[ x = [y, \dot{y}, a_i, a_h] \]  \hspace{1cm} (11)

Where, \( y \) is the relative distance at the axis \( z \), \( a_i \) and \( a_h \) are the lateral acceleration of pursuer and escaper. The pursuer uses proportional guidance:

\[ \dot{a_i} = Nq \dot{v}_e = N \frac{f_{\text{eg}} \dot{y} + \dot{y}}{v_i \dot{r}_{go}} v_e \]  \hspace{1cm} (12)

Where, \( \dot{a_i} \) is the command lateral acceleration of the pursuer, \( q \) is the LOS angle, \( v_e \) is the relative speed. Assume that both the pursuer and the high-speed aircraft are first-order time delays:

\[ \frac{a_i}{\dot{a_i}} = \frac{1}{\tau_i s + 1}, \quad \frac{a_h}{\dot{a_h}} = \frac{1}{\tau_h s + 1} \]  \hspace{1cm} (13)

Where, \( \dot{a_h} \) is the command lateral acceleration of escaper, \( \tau_i \) and \( \tau_h \) are time constants. The corresponding established state space is:

\[ \begin{cases}
\dot{x} = Ax + Bu \\
y = c^T x
\end{cases} \]  \hspace{1cm} (14)

Where,

\[
A = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & -1 & 1 \\
0 & \frac{N_{Ia}}{T_{go}} & \frac{N_{Ia}}{T_{go}} & -\frac{1}{\tau_i} & 0 \\
0 & 0 & 0 & -\frac{1}{\tau_h}
\end{bmatrix}
\]  \hspace{1cm} (15)

\[
b = \begin{bmatrix}
0 \\
0 \\
0 \\
1/\tau_h
\end{bmatrix}, u = \dot{a_h}, c = \begin{bmatrix} 1 \ 0 \ 0 \ 0 \end{bmatrix}
\]

The miss distance expression of this model is:

\[ z(t) = c^T \phi(t_f, t)x + \int_t^{t_f} c^T \phi(t_f, \tau) Bu d\tau \]  \hspace{1cm} (16)

Where, \( z(t) \) is the miss distance, \( \phi(t_f, t) \) is the state transition matrix. For SMD, its objective function is:
\[ J = \int_{t_0}^{t_f} u^2 \, dt \]  
\[ \text{s.t.} \|y(t)\| \geq M \]  

Make \( Y(t) = \phi(t, \tau) \) \( c = [e_x, e_y, e_z] \). According to (16), we can get:

\[ z_i(t_j) = c^T \phi(t_j, t_0) x_0 + \int_{t_i}^{t_j} c^T \phi(t_j, \tau) b d \tau \]
\[ = Y^T(t_j - t_0) x_0 + \int_{t_i}^{t_j} E_i(t_j - \tau) / \tau_{ud} d \tau \]  

(18)

Assume that the high-speed aircraft has a maximum overload constraint \( |u| \leq u_{\text{max}} \), then the max miss distance and the corresponding control are:

\[ \| z_i(t_j) \| = \text{sign}(Y^T(t_j - t_0) x_0) Y^T(t_j - t_0) x_0 \]
\[ + \frac{u_{\text{max}}}{\tau_{ud}} \int_{t_i}^{t_j} |E_i(t_j - \tau)| d \tau \]  

(19)

\[ u(t) = \text{sign}(Y^T(t_j - t_0) x_0) E_i(t_j - \tau) u_{\text{max}} \]  

(20)

For SMD, by setting \( \| z_i(t_j) \| = M \) we can be obtained:

\[ u_{\text{max}} = \tau_{ud} \left( M - \text{sign}(Y^T(t_j - t_0) x_0) Y^T(t_j - t_0) x_0 \right) \]
\[ \int_{t_i}^{t_j} |E_i(t_j - \tau)| d \tau \]  

(21)

Since the value \( Y(t) \) is related to the time to fly \( t_\text{go} \), the estimation of \( t_\text{go} \) and the miss distance together affect the value and direction of the control.

But in the case of considering the pursuer’s overload constraint, the whole process will become nonlinear. The saturation time of the pursuer is difficult to analyze theoretically. Morphing control will also affect the lift coefficient, which can affect the value of the limit overload in each channel. It further increases the nonlinearity of the model. In summary, there is a relatively large solution difficulty. We should still consider implementing the intelligent algorithm mentioned above.

4. SIMULATION AND ANALYSIS

Assumes that the pursuer heads on the escaper. The initial positions of escaper and the pursuer are \([100\text{km},30\text{km},100\text{km}]\) and \([0\text{km},2\text{km},0\text{km}]\). The initial speeds of the escaper and the pursuer are \(V_{\text{go}} = 2500\text{m/s} \) and \(V_{\text{go}} = 2500\text{m/s} \) respectively. The initial ballistic inclination angles of the two sides are \(\theta_{\text{go}} = -0.2\) and \(\theta_{\text{go}} = 25\). The heading angles are -130° and 35° respectively. The overload limits are both 3.5g. We set the capture radius as 15m. As for step size, variable-step is adopted: when the distance between them is greater than 20km, the step is 0.01s, when it is less than 5km, the step is 0.0001s. Accordingly, the control step size of the pursuer is the same as the simulation step, and the decision step of the escaper is 0.1s constantly.

| TABLE 2. THE INITIAL CONDITIONS OF BOTH SIDES IN THE CONFRONT |
|---|---|---|---|---|---|
| Escaper | Pursuer |
| x/km | 100 | 0 | | | |
| z/km | 100 | 0 | | | |
| y/km | 30 | 2 | | | |
| \(\theta^o\) | -0.2 | 25 | | | |
| \(\varphi^o\) | -130 | 35 | | | |
| \(V^o/(m/s)\) | 2500 | 2500 | | | |

The high-speed aircraft mainly adopts the BTT control mode. As a result, the actions of escaper mainly include angle of attack, bank angle, and morphing control, which can be expressed as \(c = [\alpha, \varphi, \kappa] \). There are certain restrictions on the angle of attack and the bank angle, leading to ranges of \(\alpha \in [-10, 20] \), \(\varphi \in [-90, 90] \) and \(\kappa \in [0,1] \). Moreover, the rate of angular change is limited to less than 10°/s.
To reduce the search space and improve the training convergence speed, the angle of attack and the bank angle are discretized as:

\[ \alpha \in [-10', -2.5', 5', 12.5', 20'] \]
\[ \phi \in [-90', -67.5', -45', -22.5', 0', 22.5', 45', 67.5', 90'] \] (22)
\[ \kappa \in [0, 0.5, 1] \]

The state space contains elements below: the relative distance \( R \) between the two sides, the planar LOS angle \( q_\beta \) at the vertical, the planar LOS angle \( q_\epsilon \) at the horizontal, the escaper velocity \( V \), the inclination angle of velocity \( \theta \). As a result, the state space can be expressed in vector form:

\[ [R, q_\beta, q_\epsilon, V, \theta] \] (23)

The design of the reward function is significant in the reinforcement learning algorithm. The reward function plays a decisive role in not only the convergence of the training but also the quality of the training result. We propose a reward function: when the miss distance is less than 15m, a relatively large negative reward -10 is given, punishing the defeat; at the final time, the remaining energy of the high-speed aircraft is used as the reward value with some processes, as shown in (25).

\[
R_D = \begin{cases} 
-10, & r_{\text{miss}} \leq r_0 \\ 
0, & r_{\text{miss}} > r_0 
\end{cases} 
\] (24)

\[ R_E = e^z - 6 \]

\[ \zeta = \left(\frac{1}{2}v^2 + gh\right)/1e6 \] (25)

The neural network structure is a fully connected neural network with 2 hidden layers, the number of hidden layer units is (256, 256), and the relu activation function is used. The network learning rate is set to 0.001, the experience pool capacity is 2000, and the sample data size is 256.

4.1 Evasion Strategy Generations

Figure 1. Learning curves.

Figure 2. Command control curves of AOA, bank angle, and morphing degree.
From Fig. 1, the training converges after about 1500 episodes. In the test of the trained model, the reward value is 22.7, greater than 20.41, which is obtained by only optimizing the angle of attack and the angle of inclination, indicating that the morphing control can improve the effectiveness of control. At the same time, the 155.2m miss distance meets the requirement.

Observing Fig. 2, it can be found that the bank angle is always -67.5° in the first 28s, and it starts to decrease to -40° after 28s. The attack angle decreases from 5° to -10°, and the morphing part also shrinks from full expansion to zero. From Fig. 3, the escaper’s overloads of \( n_y \) and \( n_z \) have both changed, resulting in the pursuer reaching saturation of \( n_z \), and the overload of \( n_y \) increased. The process of the overload change starts from 28s, which is close to the encounter time. But this strategy confuses designers because of the “black box” process, leading to low confidence in the result. We take theoretical analysis to make it better understood and more reliable.

4.2 Strategy Comprehension

For the linearized model, \( \tau_l \) and \( \tau_{li} \) are both 0.5s. Initial distance and leading angles are consistent with the nonlinear model, 150km, 180°, and 0° respectively. With the constant maximum overload (21), the final miss distance is 6m, which is lower than the requirement. Meanwhile, with the maneuver strategy based on linearized SMD, the miss distance meets the requirement, as shown in Fig. 4.
From Fig. 2 and Fig. 5, the switching of the acceleration direction is basically the same as that under the linear model. Comparing Fig. 3 and Fig. 5, it shows that under the nonlinear model, due to the constraint of the pursuer’s overload and the effect of the pitch channel, the overload generated by the high-speed aircraft is smaller, compared with the linear model. The absolute value is less than 1g, but a larger miss amount is produced. Therefore, for nonlinear problems, if the overload value is solved by (32), and the acceleration switching direction is determined by (31), the miss distance requirement can be completely satisfied. Correspondingly, the evasion strategy under the nonlinear model is somewhat comprehensible, which is mainly determined by the $t_{\text{pr}}$. That is to say, the MC-DQN method has learned the mechanism through states (23) for estimating $t_{\text{pr}}$ accurately. Simultaneously, strategy generated by training considers more on the three-dimension through $q_{\beta}$ and $q_{\gamma}$.

5. CONCLUSION
This paper trains the evasion strategy by MC-DQN under the nonlinear model. Based on the theoretical analysis of the linearization method with the objective of SMD, we obtain the interpretable evasion strategy under the nonlinear model. After that, the following conclusions are drawn:

(1) This paper analyzes the characteristics of the PE game problem and adopts the MC-DQN algorithm to solve the evasion strategy of high-speed aircraft. It proves that MC-DQN is effective for this problem.

(2) Under the nonlinear model, the change time of overload is similar to that under the nonlinear model. Therefore, the change of overload symbol (20) under the linear model can be used to interpret and apply the evasion strategy. That is, the generation of evasion strategy can be realized by accurately predicting the $t_{\text{pr}}$ in real time. The end-to-end reinforcement learning makes it successful.

The Monte-Carlo based deep reinforcement learning algorithm can fully consider the nonlinear characteristics and complex constraints of the model to generate end-to-end evasion strategy. Through linearized analysis, we make the strategy more comprehensible and reliable, which provides a new idea to explain reinforcement learning.

REFERENCES
[1] J. Shinar and D. Steinberg, “Analysis of optimal evasive maneuvers based on a linearized two-dimensional kinematic model,” Journal of Aircraft, vol. 14, pp. 85-90, 2012.
[2] G. L. Slater and W. R. Wells, “Optimal evasive tactics against a proportional navigation missile with time delay,” Journal of Spacecraft and Rockets, vol. 10, pp. 309-313, 1973.
[3] Y. Q. Wang, G. D. Ning, X. F. Wang, M. R. Hao and J. H. Wang, “Maneuver penetration strategy of near space vehicle based on differential game,” Acta Aeronautica et Astronautica Sinica, vol. 41, pp. 724, 2020.
[4] S. Y. Ong and B. L. Pierson, “Optimal planar evasive aircraft maneuvers against proportional navigation missiles,” Journal of Guidance, Control, and Dynamics, vol. 19, pp. 1210-1215, 1996.
[5] M. Pontani and B. A. Conway, “Optimal interception of evasive missile warheads: Numerical solution of the differential game,” Journal of guidance, control, and dynamics, vol. 31, pp. 1111-1122, 2008.

[6] Z. Chi, C. Li and J. Wang, “Analysis of Optimal Evasive Maneuvers Control in Near-sapce,” The 2nd China Aerospace Safety Conference, 2017.

[7] J. Ben-Asher, E. M. Cliff and H. J. Kelley, “Optimal evasion with a path-angle constraint and against two pursuers,” Journal of Guidance, Control, and Dynamics, vol. 11, pp. 300-304, 1988.

[8] J. Karelahti, K. Virtanen and T. Raivio, “Near-optimal missile avoidance trajectories via receding horizon control,” Journal of Guidance, Control, and Dynamics, vol. 30, pp. 1287-1298, 2007.

[9] M. Weiss and T. Shima, “Minimum effort pursuit/evasion guidance with specified miss distance,” Journal of Guidance, Control, and Dynamics, vol. 39, pp. 1069-1079, 2016.

[10] D. Silver, A. Huang and C. J. Maddison, “Mastering the game of Go with deep neural networks and tree search,” nature, vol. 529, pp. 484-489, 2016.

[11] J. L. Zuo, R. N. Yang, Y. Zhang, Z. L. Li and M. Wu, “Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning,” Acta Aeronautica et Astronautica Sinica, vol. 38, pp. 321168, 2017.

[12] B. Vlahov, E. Squires and L. Strickland, “On developing a uav pursuit-evasion policy using reinforcement learning,” IEEE 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 859-864, 2018.

[13] J. Ortega, N. Shaker and J. Togelius, “Imitating human playing styles in super mario bros. Entertainment Computing,” Entertainment Computing, vol. 4, pp. 93-104, 2013.

[14] J. L. Li, W. C. Chen and C. W. Min, “Terminal hypersonic trajectory modeling and optimization for maneuvering penetration and precision strike,” Journal of Beijing University of Aeronautics and Astronautics, vol. 44, pp. 556-567, 2018.

[15] B. Gaudet and R. Furfaro, “Terminal Adaptive Guidance for Autonomous Hypersonic Strike Weapons via Reinforcement Learning,” arXiv preprint arXiv:2110.00634, 2021.

[16] B. Gaudet, K. Drozd and R. Furfaro, “Adaptive Approach Phase Guidance for a Hypersonic Glider via Reinforcement Meta Learning,” AIAA SCITECH 2022 Forum, 2214, 2022.

[17] C. Liu, C. Dong and Z. Zhou, “Barrier Lyapunov function based reinforcement learning control for air-breathing hypersonic vehicle with variable geometry inlet,” Aerospace Science and Technology, vol. 96, pp. 105537, 2020.

[18] B. Gaudet, R. Linares and R. Furfaro, “Deep reinforcement learning for six degree-of-freedom planetary landing,” Advances in Space Research, vol. 65, pp. 1723-1741, 2020.

[19] Q. H. Zhou, Y. F. Liu, N. M. Qi and J. F. Yan, “Anti-warning-based anti-interception avoiding penetration strategy in midcourse,” Acta Aeronautica et Astronautica Sinica, vol. 38, pp. 319922, 2017.

[20] Y. X. Mei, “Fast path planning and longitudinal guidance of hypersonic reentry vehicles with multiple constraints,” Xiamen University, 2019.

[21] N. Dhananjay and D. Ghose, “Accurate time-to-go estimation for proportional navigation guidance,” Journal of Guidance, Control, and Dynamics, vol. 37, pp. 1378-1383, 2014.

[22] D. Pathak, P. Agrawal and A. A. Efros, “Curiosity-driven exploration by self-supervised prediction,” International conference on machine learning, PMLR, pp. 2778-2787, 2017.

[23] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," arXiv preprint arXiv:1511.05952, 2015.

[24] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," arXiv preprint arXiv:1707.01495, 2017.

[25] J. Shinar, Y. Rotszteim and E. Bezner, “Analysis of three-dimensional optimal evasion with linearized kinematics,” Journal of Guidance, Control, and Dynamics, vol. 2, pp. 353-360, 1979.
[26] Guelman, “On the Divergence of Proportional Navigation Homing System,” in Armament Development Authority Israel Ministry of Defense, 1969.

[27] B. Yang, Y. C. Zhu, Y. M. Wei and Z. C. Fan, “Trajectory optimization and control analysis of folding wing aircraft,” Chinese Space Science and Technology, vol. 40, pp. 64-75, 2020.