Autonomous decision system for UAV control in 3D simulated environment using deep neural networks

B. FIRLEJ, P. SUJECKI

bfirlej@mon.gov.pl, przemo.sujecki@gmail.com

Military University of Technology, Faculty of Cybernetics
Kaliskiego Str. 2, 00-908 Warsaw, Poland

This article presents an autonomous decision-making system for unmanned aerial vehicles (UAVs) operating in GPS-denied environments. The proposed solution is based on reinforcement learning and is intended to support mission execution when external positioning signals and continuous operator guidance are unavailable or unreliable. The system is implemented in a Unity-based 3D simulation environment and employs a PPO-trained Actor–Critic neural architecture with continuous control outputs. The primary operational scenario considered in this work is the autonomous protection of a mothership drone, where the UAV must maintain effective defensive positioning, preserve mission continuity, and react to dynamic threats and environmental constraints. In addition, a strike mission is included as a secondary task configuration to demonstrate the adaptability of the proposed framework to different UAV operational roles. The article discusses the conceptual structure of the system, the observation design, the control formulation, and the role of reward engineering in shaping autonomous mission behavior.

Keywords: deep neural networks, UAV, autonomous.

DOI: 10.5604/01.3001.0055.7727

1. Introduction and research motivation

Unmanned aerial vehicles (UAVs) are increasingly employed in autonomous missions such as inspection, environmental monitoring, and search-and-rescue operations. Despite this progress, the majority of existing UAV navigation systems remain strongly dependent on Global Positioning System (GPS) signals and external communication links for localization and waypoint tracking. These mechanisms become unreliable or unavailable in environments affected by electromagnetic interference, urban canyons, dense vegetation, or intentional signal disruption, including GPS jamming and spoofing attacks [1], [2]. Consequently, traditional navigation and control architectures are insufficient for ensuring robust UAV operation in contested or GPS-denied environments. To overcome these limitations, recent research has increasingly focused on artificial intelligence – based approaches aimed at reducing reliance on external positioning infrastructure. Vision-based navigation systems combined with deep learning techniques enable UAVs to extract spatial and temporal information directly from onboard sensors. In particular, convolutional neural networks (CNNs) and deep reinforcement learning (DRL) methods have demonstrated the ability to perform perception-driven navigation without explicit global localization [3]. Moreover, integrated deep learning and reinforcement learning control models allow UAVs to maintain situational awareness and adaptive behavior under uncertain environmental conditions [4].

Among AI-based control methods, reinforcement learning (RL) has attracted significant attention due to its ability to learn optimal control policies through interaction with the environment. Deep RL algorithms, such as DQN, DDPG, and PPO, have been successfully applied to UAV tasks including obstacle avoidance, trajectory optimization, and autonomous path planning [1], [5], [6]. Actor--Critic architectures are particularly well suited for UAV applications, as they effectively address continuous control problems related to flight dynamics, attitude stabilization, and smooth trajectory execution [3]. Comparative studies indicate that RL-based controllers often outperform classical control approaches, such as PID and model predictive control (MPC), in terms of adaptability and robustness to environmental disturbances [6].

Despite promising results, current learning--based UAV control approaches still face several

unresolved challenges. Many proposed methods assume the availability of auxiliary navigation cues or partial localization signals, which implicitly reintroduce dependence on external systems and limit full autonomy in GPS-denied scenarios [7]. Additionally, deep reinforcement learning techniques are associated with high sample complexity, limited generalization from simulation to real-world environments, and reduced interpretability of learned policies, which hinders their deployment in safety-critical UAV applications [8], [9].

In contrast to existing approaches discussed in the literature, this work proposes an autonomous UAV decision-making system based on a custom-designed Actor-Critic neural network trained using the Proximal Policy Optimization (PPO) algorithm. Unlike many prior studies that focus primarily on target--reaching or trajectory-following tasks, the proposed solution is structured as a mission--oriented control framework that can be adapted to different operational roles.

The learning process is conducted within an author-developed simulation environment implemented in Unity, which enables full control over mission conditions, environmental structure, and observation design.

The primary objective of this study is to develop an autonomous UAV decision-making system for mothership protection in a GPS-denied 3D simulated environment. The system is intended to support autonomous operation under degraded navigation conditions by enabling the UAV to maintain a protective role around a higher-value platform, react to threats, and preserve mission continuity without relying on waypoint-based navigation or external positioning signals. In addition, a strike mission scenario is included as a secondary task configuration in order to demonstrate the flexibility of the proposed reinforcement learning framework.

The main contributions of this work are as follows:

– the development of an author-designed 3D Unity simulation environment for UAV training in GPS-denied conditions;
– the design of a custom PPO-based Actor--Critic control architecture tailored to UAV sensory inputs and continuous control;
– the formulation of a reward structure for autonomous mothership protection, including spatial positioning, threat interception, safety constraints, and mission continuity;
– the demonstration that a single reinforcement learning framework can be adapted to different UAV mission types through task-specific reward design.

2. AI Agent based on Reinforcement Learning

2.1. Concept and role in UAV control

In the proposed approach, the control of an unmanned aerial vehicle (UAV) is delegated to an artificial intelligence (AI) agent trained with reinforcement learning (RL). Instead of following a fixed, pre-programmed procedure, the UAV learns a control policy through trial and error in a simulated environment.

Building on the proposed approach, it is important to note that UAVs often operate in environments where GPS signals may be weak, jammed, or completely unavailable – such as in urban canyons, dense forests, or military scenarios. Therefore, the integration of an AI--based control system becomes essential.

By leveraging reinforcement learning, the UAV can maintain autonomous, ensures robust navigation, stability, and mission continuity despite signal loss, effectively allowing the UAV to operate independently and safely even in the absence of GPS signals. The AI agent becomes a decision-making module that continuously observes the state of the environment and selects actions such as changes of thrust, roll, pitch and yaw, or higher-level maneuvers (e.g. evasive action, approach to target, orbiting).

In the primary mission scenario considered in this work, target approach is interpreted as an interception-oriented behavior directed against a hostile drone threatening the mothership. The observation space consisted of 128 input variables. This dimensionality was selected from an engineering perspective in order to provide the agent with a sufficiently complete representation of its internal flight state and available sensor--related information.

The observation vector was designed to include the data necessary for stable control, obstacle avoidance, and target-approach behavior.

In the primary protection scenario, this behavior corresponds to approaching and intercepting a hostile aerial target in order to defend the mothership.

No separate study was conducted to determine the minimal or globally optimal observation dimensionality; instead, the

observation set was defined as a practical design choice for the considered mission scenario.

Tab. 1. Selected onboard UAV observation categories

Sensor	**Function**
Position	Determines UAV location within the environment
Velocity	Describes UAV movement dynamics
Orientation (roll, pitch, yaw)	Defines UAV attitude and heading
Angular rates	Provides rotational motion information
Acceleration	Supports motion estimation and dynamic response
Attitude estimates	Enables stable orientation control
Relative motion	Supports navigation and trajectory adjustment

From the perspective of control architecture, the RL agent replaces or augments classical guidance and navigation process. It does not need an explicit mathematical model. Instead, it learns directly from environment.

The training environment was developed in Unity 3D and consisted of a UAV agent, a protected mothership platform, hostile aerial targets, and obstacles distributed within the operational area. The environment was designed primarily to reproduce a mothership-protection scenario in which the UAV had to maintain effective defensive positioning, approach and intercept hostile drone when necessary, avoid collisions, and preserve mission continuity without relying on GPS or waypoint-based navigation. In each state, the agent takes its current observations and processes them to produce an action in response to these observations, in a way that maximizes long-term mission performance as expressed by a reward function. This makes the approach particularly attractive in complex, uncertain and adversarial scenarios typical for military operations, where hand-crafted control logic may be incomplete.

In the present work, the proposed decision--making framework is analyzed in two mission configurations.

The primary configuration is mothership protection, in which the UAV is required to defend a higher-value platform against hostile aerial threats. The secondary configuration is a strike mission, included to demonstrate that the same reinforcement learning framework can be adapted to a different operational objective through task-specific reward design.

2.2. Neural network policy and representation of the Agent

At the core of the RL agent lies a deep neural network that represents its policy. The policy is a function that receives a vector of observations describing part of the current state and outputs a probability distribution over possible actions.

Observations for a quadrotor UAV include:

- Relevant data provided by the flight controller such as position, velocity, orientation, angular rates, battery level;
- Sensor readings: distances to obstacles (e.g. simulated lidar or depth sensors), radar/EO detections, relative position of threats and targets;
- Mission context: current waypoint, time remaining, rules of engagement flags, threat level indicators.

Based on the combination of these inputs, the drone can execute a variety of autonomous and adaptive behaviors. For instance, using position, velocity, and orientation data, the agent can maintain stable flight, perform trajectory tracking, and execute precise waypoint navigation. Angular rates and thrust control allow the UAV to handle disturbances such as wind or payload shifts by applying corrective maneuvers in real time.

With lidar or depth sensor inputs, the agent can detect and avoid obstacles dynamically, even in cluttered or GPS-denied environments. EO/IR or radar detections enable the UAV to identify and track moving targets or threats, adjust flight paths for surveillance or reconnaissance, and maintain optimal sensor line-of-sight. Relative position and velocity estimates of other aerial or ground assets allow for cooperative multi-agent coordination, such as swarm formation, collision avoidance, and task allocation among multiple drones.

The mission context inputs – including current waypoint, time constraints, and engagement rules – guide the agent's high-level decision-making. For example, the UAV can prioritize safe navigation in low-battery conditions, switch between surveillance and engagement modes based on threat indicators, or dynamically re-plan routes when mission objectives change.

The network typically consists of multiple fully connected layers, sometimes combined with recurrent layers (for partially observable

scenarios where memory is needed). The parameters (weights and biases) of this network are not manually designed; they are learned during training through interaction with the environment.

In current more advanced configuration, the agent uses an Actor-Critic architecture, where:

- the Actor network outputs the policy (which action to take),
- the Critic network estimates the value function (how good a given state or state–action pair is in terms of expected future reward).

This separation allows for more stable and sample-efficient training compared to basic policy-gradient methods. The network architecture was designed from an engineering perspective and iteratively adjusted during development to obtain stable learning behavior and satisfactory control performance. The number of layers and neurons was not determined through a separate formal optimization study reported in this paper. Instead, the architecture was refined pragmatically based on training behavior and implementation experience.

2.3. Reinforcement learning and the role of the reward function

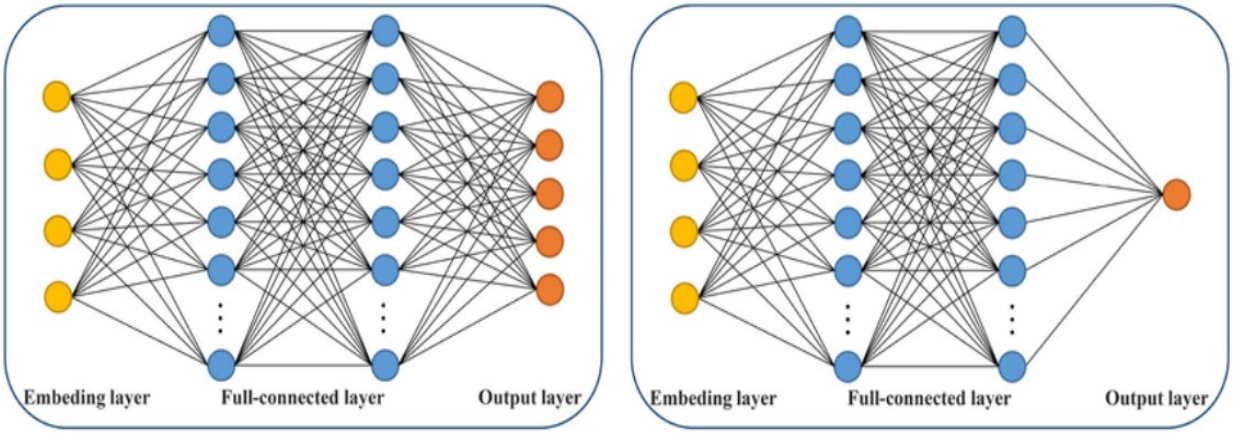
Reinforcement learning formalizes the interaction between the UAV (agent) and its environment as a Markov decision process. At each time step:

1. The agent receives an observation of the current state.
2. It selects an action according to its policy.
3. The environment responds with a new state and a scalar reward signal.
4. The agent updates its policy parameters to maximize the expected cumulative reward over time.

The reward function is the central mechanism by which mission objectives and operational constraints are encoded. It translates high-level goals into a numerical signal that the agent can optimize.

For UAV control in military-type missions, the reward function can combine several components:

1. Task completion:
   - Positive reward for reaching the target area, maintaining coverage of a protected zone, or successfully intercepting an incoming threat.
   - Larger terminal reward for fully accomplishing the mission (e.g. “hit target”, “mothership remains unharmed until mission end”).
2. Survivability and safety:
   - Negative rewards (penalties) for collisions with terrain, buildings, friendly units or obstacles.
   - Penalties for entering prohibited zones, violating minimum separation distances, or exposing the UAV to excessive risk (e.g. flying within known enemy engagement envelopes).
3. Efficiency and timeliness:
   - Small negative reward proportional to mission duration or energy consumption, encouraging the agent to complete its task faster and with fewer maneuvers.
   - Penalties for unnecessarily aggressive maneuvers that increase detectability.



(a) Actor network (b) Critic network

Fig. 1. Architecture of neural networks used in configuration
Source: https://www.researchgate.net/ figure/Structures-of-the-Actor-network-and-Critic-network-in-the-MAR-PPO-framework_fig4_385822792

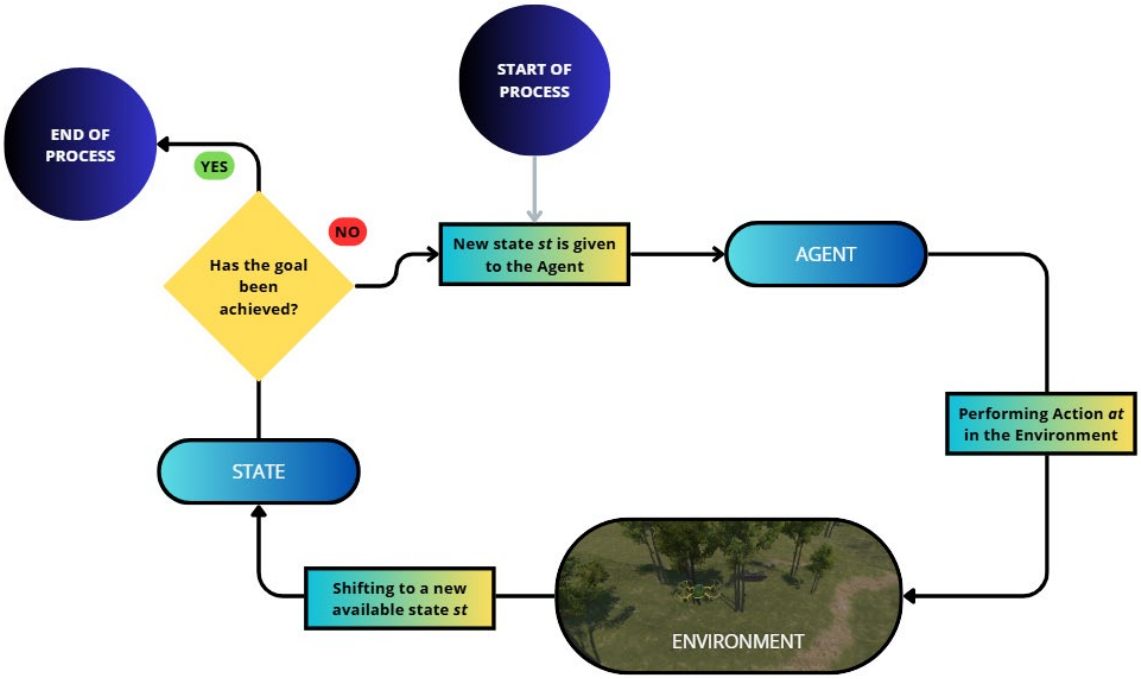


Fig. 2. Process of training AI Agent

By carefully balancing these components, one can guide the agent towards desired behavior: it should not only reach the target, but also do so safely, efficiently and within tactical constraints. The design of the reward function is therefore a crucial engineering step, especially in safety--critical contexts.

3. Mission Scenarios: Mothership Protection and Strike Mission

In the present work, two mission scenarios are considered within the same reinforcement learning framework. The primary scenario is mothership protection, while the strike mission is treated as a secondary task configuration demonstrating the adaptability of the proposed autonomous decision system.

3.1. Protection of a "Mothership" Drone

In a protective role, the reinforcement learning (RL) agent functions as an escort or "shield" for a more valuable mothership UAV. Its primary mission is to ensure the survival of the mothership by actively defending against incoming threats and maintaining optimal positioning relative to both the mothership and adversaries.

Tab. 2. Design rewards for protection mission

Reward Component	Description	Type
Mothership Survival	The agent receives the highest reward if the mothership remains undamaged until the end of the mission episode, reflecting successful protection.	Positive
Threat Interception	Intermediate rewards are given for intercepting or neutralizing hostile UAVs, missiles, or projectiles before they reach the mothership.	Positive
Disruption of Enemy Attack	Reward for forcing threats to disengage, lose lock, or adopt unfavorable engagement geometries through evasive or diversionary maneuvers.	Positive

Protection Zone Compliance	Penalty for leaving the defined protection area around the mothership, ensuring the agent stays within effective defensive range.	Negative
Threat Response Failure	Penalty for delayed or inadequate reaction to detected threats, representing lapses in situational awareness.	Negative
Collision Avoidance	Penalty for collisions or near misses with the mothership or other friendly units, enforcing safe separation during defensive actions.	Negative

The reward function can be designed so that:

- The highest reward is obtained when the mothership remains unharmed until the end of the episode;
- Intermediate rewards are given for successfully intercepting threats or forcing them to break off;
- Penalties are given if the agent leaves the protection zone, fails to react to threats, or causes friendly collisions;

The agent learns effective interception and shielding maneuvers, balancing its own survivability with the primary goal of protecting the mothership.

3.2. Strike Mission

In a strike mission, the agent's objective is to reach and engage a designated target while surviving enemy defenses. The reinforcement learning agent is trained to navigate through complex environments, avoid threats, and select optimal attack vectors that minimize exposure while maximizing mission success probability.
The agent is trained to:

- Navigate through complex terrain or urban environments;
- Avoid known and unknown threats (e.g. simulated air defenses, hostile UAVs);
- Choose attack vectors that minimize exposure and maximize probability of mission success;

Tab. 3. Designed rewards for strike mission

Training Objective/ Reward Component	**Description**	**Type**
Approach Target	The agent receives a positive reward for each time step in which it reduces the distance to the designated impact point. This encourages continuous progress toward the target.	Positive
Deviation from Target	A penalty is applied when the UAV increases its distance from the target, discouraging inefficient or incorrect maneuvers.	Negative
Mission Complete Time	A positive reward proportional to the speed of mission completion – faster engagement leads to higher reward values.	Positive
Operational Zone Compliance	The agent is penalized for leaving the predefined mission area or crossing operational boundaries.	Negative
Flight Efficiency	Penalty for unnecessary, excessive, or oscillatory maneuvers (e.g., frequent sharp turns or altitude changes without tactical need).	Negative
Stable Flight Behavior	Additional reward for maintaining a smooth trajectory, minimizing abrupt attitude or velocity changes, and conserving energy.	Positive

Through training on diverse scenarios, the agent learns robust tactics that are not explicitly programmed but emerge from the optimization of the reward function. The reward function was iteratively tuned and refined to maximize the agent's performance and encourage stable, high-quality behaviors.

4. Advantages and Challenges

The use of reinforcement learning (RL) – based AI agents for UAV control in military operations offers key benefits. Such agents adapt to complex, dynamic and partially unknown environments without manually encoding all situations. The designer defines high-level mission objectives and constraints, and the agent learns how to react to changing battlefield conditions. After training, the agent can make real-time decisions with a high degree of autonomy, which is crucial under limited or delayed communication.

The operator focuses on defining tasks and rules of engagement instead of manually piloting the UAV. The same RL framework can be reused across different mission types by mainly changing the reward function and observation set.

Simulation enables safe and intensive training in diverse, high-risk scenarios. A key aspect are emergent behaviors: complex tactics not explicitly programmed but arising from the interaction of reward, environment and learning. They may lead to innovative maneuvers and cooperation, but can also be unsafe or misaligned with operational and ethical constraints if they exploit flaws in the reward design.

This highlights the need for careful reward shaping, safety constraints and human oversight, as well as addressing sim-to-real transfer and limited policy interpretability.

A limitation of the present study is that neither the observation dimensionality nor the neural architecture was subjected to a separate ablation study aimed at identifying an optimal configuration. These elements were established as engineering design choices for the considered mission scenarios, while the PPO training hyperparameters were refined separately.

This study presented a PPO-based Actor–Critic decision system for autonomous UAV control in a GPS-denied Unity-based simulation environment, with primary emphasis on mothership protection. The proposed framework was formulated as a mission-oriented autonomous control system rather than solely as a target-reaching policy. The analysis showed that reinforcement learning can support protection-oriented behavior by combining engineered observations, continuous control actions, and reward-driven adaptation. In addition, the inclusion of a secondary strike scenario demonstrated that the same framework can be adapted to different operational objectives through task-specific reward design.

5. Bibliography

[1] Chijioke C.E., Elfouly T., Alouani A., Khattab T., “A Survey on UAV Control with Multi-Agent Reinforcement Learning”, *Drones*, Vol. 9, No. 7, 484, Jul. 2025, DOI: 10.3390/drones9070484.

[2] Cwojdziński L., “Wpływ sztucznej inteligencji na rozwój nowoczesnych systemów bezzałogowych”, *Kwartalnik Bellona*, Vol. 719, No. 4, 43–56 (2025), DOI: 10.5604/01.3001.0055.0552.

[3] Wang J., Yu Z., Zhou D., Shi J., Deng R., “Vision-Based Deep Reinforcement Learning of Unmanned Aerial Vehicle (UAV) Autonomous Navigation Using Privileged Information”, *Drones*, Vol. 8, No. 12, 782 (2024), DOI: 10.3390/drones8120782.

[4] Liu S., “Research on UAV Intelligent Control Model Based on Deep Learning and Reinforcement Learning”, *JCSAI*, Vol. 2, No. 2, 61–64 (2025), DOI: 10.54097/49bx3p61.

[5] Sharma G., Jain S., “Deep Reinforcement Learning-Based Framework for Path Planning of AUAVs”, *Procedia Computer Science*, Vol. 258, 1112–1122 (2025), DOI: 10.1016/j.procs.2025.04.346.

[6] Miera P., Szolc H., Kryjak T., “Control of an Autonomous Unmanned Aerial Vehicle Using Reinforcement Learning”, No. 4, 85–91 (2023), DOI: 10.14313/PAR_250/85.

[7] Yang C., “Review of Ai-Based Uav Navigation in Gps-Denied Environments”, *Proceedings of the 2025 2nd International Conference on Electrical Engineering and Intelligent Control* (EEIC 2025), Advances in Engineering Research, Vol. 279, pp. 583–596, Atlantis Press International BV, Dordrecht 2025, DOI: 10.2991/978-94--6463--864-6_51.

[8] Khattak W.R., Asad M., Ahmad W., “Deep reinforcement learning in UAV flight control and navigation: A systematic review of algorithms, benchmarks, and safety”, *Spectrum of Engineering Sciences*, Vol. 4, No. 1, 806–820 (2026).

[9] Kaufmann E., Bauersfeld L., Loquercio A., Müller M., Koltun V., Scaramuzza D., “Champion-level drone racing using deep reinforcement learning”, *Nature*, Vol. 620, No. 7976, 982–987 (2023), DOI: 10.1038/s41586-023-06419-4.

Autonomiczny system decyzyjny do sterowania bezzałogowymi statkami powietrznymi w symulowanym środowisku 3D z wykorzystaniem głębokich sieci neuronowych

B. FIRLEJ, P. SUJECKI

W niniejszym artykule przedstawiono autonomiczny system podejmowania decyzji dla bezzałogowych statków powietrznych (UAV) działających w środowiskach pozbawionych sygnału GPS. Proponowane rozwiązanie opiera się na uczeniu przez wzmacnianie i ma wspierać realizację misji w przypadku niedostępności lub zawodności zewnętrznych sygnałów pozycjonujących i ciągłego sterowania operatora. System został zaimplementowany w środowisku symulacji 3D bazującym na Unity i wykorzystuje architekturę neuronową Aktor–Krytyk, wyszkoloną w PPO, z ciągłymi wyjściami sterującymi. Głównym scenariuszem operacyjnym rozważanym w tej pracy jest autonomiczna ochrona drona-matki, w której UAV musi utrzymywać skuteczne pozycjonowanie obronne, zachować ciągłość misji oraz reagować na dynamiczne zagrożenia i ograniczenia środowiskowe. Dodatkowo, misja uderzeniowa została uwzględniona jako konfiguracja zadania drugorzędnego, aby zademonstrować adaptowalność proponowanych ram do różnych ról operacyjnych UAV. W artykule omówiono strukturę koncepcyjną systemu, projekt obserwacji, formułę sterowania oraz rolę inżynierii nagród w kształtowaniu zachowania podczas misji autonomicznych.

Słowa kluczowe: głębokie sieci neuronowe, UAV, autonomiczny system.