International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-11

https://doi.org/10.70917/ijcisim-2026-4138
Article

Dynamic Optimization of Police Dog Task Allocation Based on Reinforcement Learning: An Ethical Intervention Strategy for Stress Reduction

Jinglin Li [1,*], Xiaofu Pan [1], Cuiping Gou [1] and Kexiao Liu [1]

[1] College of State Governance, Southwest University, Chongqing, 400715, China
* Correspondence author: jinglinli0708@163.com

Abstract: This paper presents a dynamic optimization model for police dog task allocation based on reinforcement learning. This model can be regarded as a Markov decision process (MDP), which formally Outlines how the environment and the dog's state are interrelated, as well as the sequence of stress accumulation and relief, and uses state vectors to dynamically adjust the allocation plan. The experimental results show that the proposed method achieves a task completion rate of 94.3%, reduces the required rest intervention by 46-57%, and is an active stress prevention measure rather than a passive threshold enforcement mechanism. Even when the estimation error of the stress-sensitive parameter was 20%, the completion rate was still 92.7%, and the average stress was 0.42. The research provides a feasible approach to balance operational requirements and ethical responsibilities in the management of working animals.

Keywords: reinforcement learning, police dog, task allocation, dynamic optimization

1. Introduction

Police dogs are required components of the new security system and have many functions, such as detecting explosives and drugs, locating criminals, searching for lost children and pets, etc. Only in the United States are there about 50,000 police dogs across all levels of law enforcement, and the cost of acquiring, training and equipping each dog exceeds $20,000. The Working Value of these animals is directly related to their physical condition, cognitive alertness and emotional stability [1]. Therefore, if a trained police dog dies or is forced to retire early due to stress-related illness, there will be considerable replacement costs and operational disruption, resulting in both economic and moral damage.

Prolonged exposure to high-intensity work leads to accumulated stress and thus reduced efficiency and health problems. The reasons for the above stress responses are known to us [2]. When the police dog finds an external trigger, such as a person carrying a weapon or an explosion that causes a bomb to go off, the HPA axis is activated, and cortisol and adrenaline are released. An acute-stress response can increase performance, but continuous or extended activation without proper recovery leads to allostatic overload; that is to say, damage to the body's systems due to cumulative effects [3]. A stress response in a working dog is shown by elevated cortisol levels, a high heart rate, behavioural inhibition and reduced olfactory accuracy. Haverbeke and others have conducted research to find that police dogs that have undergone numerous tactical deployments show elevated basal cortisol levels compared with kennel-rested control dogs; thus, it can be inferred that they have not fully recovered from their previous duties. Similarly, Darby and his colleagues found that dogs undergoing high-intensity work without individualised rest periods showed a decline in their ability to detect scent gradually, and error rates increased by as much as 23 per cent over the course of a two-week deployment [4].

The general division of tasks in police canine units is either a fixed-rotation schedule or

handler-based discretion, and neither systematically considers individual stress changes or recovery needs. The heterogeneity of individual dog stress responses is relatively large, so these static methods are not very effective [5]. Breed-based allocation rules are administratively simple but fail to consider the large differences in temperament, resilience and stress sensitivity among breeds documented by Rooney and others [6]. A Belgian Malinois selected for patrol work based solely on breed characteristics may have stress-sensitive parameters that make it unsuitable for extended high-intensity operations, while a Labrador Retriever initially designated for detection duties might exhibit unexpected stress resilience and thus be suitable for more demanding tasks. The static allocation decisions made in the initial training and certification cannot adapt to the changes in individual dogs' physiology and psychology with age and experience, or health problems. In addition, the time evolution of stress accumulation and recovery leads to a sequence of decisions that static scheduling methods are not suitable for. A dog trained for an apprehension task may need 48-72 hours to fully recover physically after a high-intensity task before being subjected to another, but fixed rotation schedules often fail to consider this individual recovery cycle [7].

Given the demands of production and animal welfare, new ways of organising work that consider these considerations and are still humane need to be explored. The new area of animal-computer interaction has started to explore technology for observing and enhancing the welfare of working animals, but few studies have addressed the specific problem of predictive task allocation under stress. Reinforcement Learning can be applied to this problem; it handles sequential decision-making in a stochastic environment and, by designing an appropriate reward function, incorporates multiple objectives. Reinforcement learning is not a traditional optimisation method for the static assignment problem; instead, it learns a policy that maps observed system states to optimal actions through interaction with the environment and adapts to individual differences [8].

A dynamic optimisation model based on reinforcement learning is presented in this paper to distribute tasks among police dogs in a way that reduces stress accumulation without affecting mission effectiveness. Each dog in the model is an independent agent, and the state of each dog is changed by task exposure, rest time and individual stress resilience parameters. Ethical interventions are introduced as soft constraints in the reward function to make welfare considerations affect the allocation decision at every time step. This work has the following three contributions. First, a formal Markov Decision Process is constructed to address the problem of police dog task allocation under stress-aware state representations. Second, an ethical reward shaping mechanism is added to quantify welfare costs and include them in the optimisation objective [9]. Third, numerous simulation experiments verify the model with the base schedule method under different operating loads. The following sections will introduce related studies, present the construction of the mathematical system and algorithm, show experimental results, and finally draw conclusions on the application and future directions of research.

2. Literature Review and Theoretical Basis

2.1. Research on Police Dog Task Allocation

Historically, the deployment of police dogs has been based on breed-specialist and handler-experienced systems rather than data-driven optimisation. German Shepherds and Belgian Malinois are used for patrols and apprehensions due to their physical strength; Labrador Retrievers and Beagles are selected for detection work because they have good sense of smell and are more docile [10]. Rooney and his colleagues have studied the performance differences among various dog breeds in work and found that the variation among individuals within a breed is often greater than that between breeds; therefore, distribution according to breed alone is not ideal. Research in the military and police has shown that, by assigning dogs to tasks based on individual behaviour, both the success rate and injury risk have been reduced. However, these matching processes are relatively fixed and only occur during the initial training and certification stages, not continuously throughout a person's working life [11].

Dynamic Task Allocation has received relatively little attention in the field of canine working animals. Research on human resources provides some analogous cases through nurse scheduling, emergency responder dispatch and military personnel rotation. The above regions use optimisation techniques such as integer programming, genetic algorithms and constraint satisfaction to balance workload fairness with skill requirements. Adapting such methods to police dog units faces particular difficulties because the workers are non-human animals, and their wishes and limitations cannot be directly communicated. Therefore, proxy indicators of well-being and stress are used instead, often in the form of physiological sensors and behavioural observation protocols.

2.2. Stress Physiology and Welfare Science

Stress in mammals activates the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system; thus, many indices of stress can be observed, such as salivary cortisol, blood glucose, heart rate variability and core body temperature. The Stress Response in working dogs is affected by genes, early-life experiences, training history and health problems. Prolonged stress weakens the immune system [12]; it can cause ulcers in the stomach and harm cognition, thus elevating the risk of death from other causes in older adults. The Five Freedoms, which were first introduced to promote the welfare of farm animals, have been adapted for working dogs to ensure that they are free from hunger and thirst, discomfort and pain, injury and disease, fear and distress, and have the freedom to behave normally.

Recently, wearable sensor technology has advanced to monitor the physical condition of dogs during exercise continuously. Accelerometers, GPS trackers, and heart-rate monitors can all acquire high-frequency data streams to monitor changes in a person's level of activity and health status during rest and stress recovery [13]. Cobb et al. have shown that heart-rate variability in wearable devices is closely related to behaviour-based stress in detection dogs. The above technologies will enable the construction of a closed-loop system that adjusts management based on physiological data.

2.3. Reinforcement Learning in Resource Allocation

Reinforcement learning has been applied to solve many problems of sequential decision-making under uncertainty in games, robotics and resource management. The first idea is that an agent learns to act reasonably in an environment through trial-and-error by receiving a reward for good behaviour and a penalty for bad behaviour. Reinforcement learning agents learn policies in the context of resource allocation problems to map system states to allocation decisions and adapt to changing demand and resource conditions over time [14].

This study can be used to schedule patients' appointments at a hospital, divide the work of a cloud computer, and manage a fleet of self-driving cars. In all cases, the agent needs to choose between immediate efficiency and the long-term health of the system. Police dog allocation has a resource of available canine workforce, demands that correspond to incoming tasks, and system health is related to cumulative stress levels and operational readiness. The problem of temporal credit assignment in reinforcement learning is consistent with the dynamics of stress accumulation; thus, the current task assignment will affect future physiological conditions and performance.

2.4. Ethical Frameworks for Animal Computer Interaction

The branch of animal-computer interaction specifically addresses the ethical problems of technology for non-human animals. Mancini proposed that the Design of such a system should be based on the animal's own needs and desires. This view does not consider animals as mere instruments for achieving human welfare under the previous utilitarianism. Ethical design of police dogs should consider individual limitations in the task-allocation algorithm and offer choices whenever feasible.

Consequentialist ethics determines the goodness of an act based on its results, and is thus suitable for optimisation-based methods. Pure consequentialism may be willing to sacrifice the interests of a few for the greater good. Deontological constraints are prohibitions on certain actions irrespective of the results, and they can supplement consequentialist optimisation by setting inviolable welfare boundaries. Both methods are employed in this study to set up deontological stress thresholds as constraints in a consequentialist reward function and thus optimise operating results within ethical bounds.

2.5. Integration of Theoretical Frameworks

The four threads of the theoretical system are: individual-based task allocation from working-dog research, stress physiology and monitoring from welfare science, sequential decision optimisation via reinforcement learning, and constraint-based ethics of animal-computer interaction. Police dog task allocation is viewed as a sequential decision problem, and the system state at each time step comprises the physiological status of all dogs, attributes of pending tasks, and environmental conditions. Actions are task assignments or rest instructions. State transitions are the physiological results of task execution and recovery. Rewards are composed of operating-success indicators and welfare-cost functions. Ethical constraints are boundary conditions that prohibit allocations with stress indicators exceeding the predefined threshold. The integrated framework will guide the development of the mathematical model in the following section.

3. Methodology and Model Construction

3.1. Problem Formulation

Consider a police canine unit comprising N dogs indexed by i in the set {1, 2, ..., N}. At each discrete time step t, a set of M tasks arrives requiring assignment. Each task j in {1, 2, ..., M} is characterized by a feature vector including task type, estimated duration, environmental intensity, and required specialization. Each dog i possesses an individual state vector $s_{i,t}$ capturing current stress level, fatigue index, recovery status, and skill proficiency. The system state S_t is the concatenation of all individual states and pending task characteristics.

The allocation decision at time t is represented by an assignment matrix A_t where element $a_{i,j,t}$ equals 1 if dog i is assigned to task j at time t, and 0 otherwise. Each dog may be assigned at most one task per time step, and some dogs may receive rest assignments represented by a special null task indexed 0. The objective is to learn a policy pi that maps system states to assignment decisions maximizing cumulative operational effectiveness while respecting individual stress constraints.

Table 1 shows the individual dog parameters used in the main simulation case. Anonymise the Dog Identifier. The range of stress sensitivity is 0.12 - 0.38, and it varies considerably among people. Recovery rates range from 0.16 to 0.33, and thus vary in the speed at which each dog returns to normal after high-demand work.

Table 1. Individual Dog Physiological Parameters

Dog ID	Stress Sensitivity	Recovery Rate	Initial Skill Level	Age Years
K01	0.18	0.28	0.82	4.2
K02	0.32	0.19	0.75	5.1
K03	0.14	0.31	0.91	3.8
K04	0.25	0.22	0.68	6.3
K05	0.38	0.16	0.79	4.9
K06	0.21	0.25	0.85	3.5
K07	0.29	0.20	0.72	5.7
K08	0.16	0.33	0.88	4.1

3.2. Markov Decision Process Formulation

The task allocation problem can be formulated as a Markov Decision Process (MDP) with the elements *{S, A, P, R, gamma}*. S is the set of all possible states for dogs' bodies and task queues. A is the action space containing all feasible assignment matrices. P is the state transition probability function. R is the reward function, and gamma in [0,1] is the discount factor that balances immediate and future rewards.

The State-Transition Probabilities are as follows:

$$P(s' \mid s, a) = prod_{i=1}^{N} P_i(s'_i \mid s_i, a_i) \tag{1}$$

where s represents the current system state, s' represents the next system state, a represents the joint action, s_i represents the state of dog i, s'_i represents the next state of dog i, and a_i represents the action affecting dog i. This factorization assumes that individual dog state transitions are conditionally independent given the assignment decision, which holds when dogs operate independently without direct behavioral contagion.

An individual's state transition probability includes stress dynamics and recovery. A dog assigned to a task will have a higher stress level if the task is demanding or if it is more sensitive. The stress of a resting dog is lower, and so on for other recovery rate parameters. The on/off is as follows:

$$s'_i = f(s_i, a_i, theta_i) + epsilon_i \tag{2}$$

where f represents the deterministic state update function, $theta_i$ represents the individual physiological parameter vector for dog i, and $epsilon_i$ represents zero mean Gaussian noise capturing unmodeled variability.

Figure 1 shows the system architecture of the proposed model. The three interconnected layers are as follows: the sensing layer includes wearable devices and behavioural observation modules; the decision layer contains the state estimator, Q-learning agent and constraint checker; and the execution layer translates allocation decisions into task assignments and rest instructions. Double-headed arrows

show that the physiological monitoring data are used to change decisions and carry out distribution.

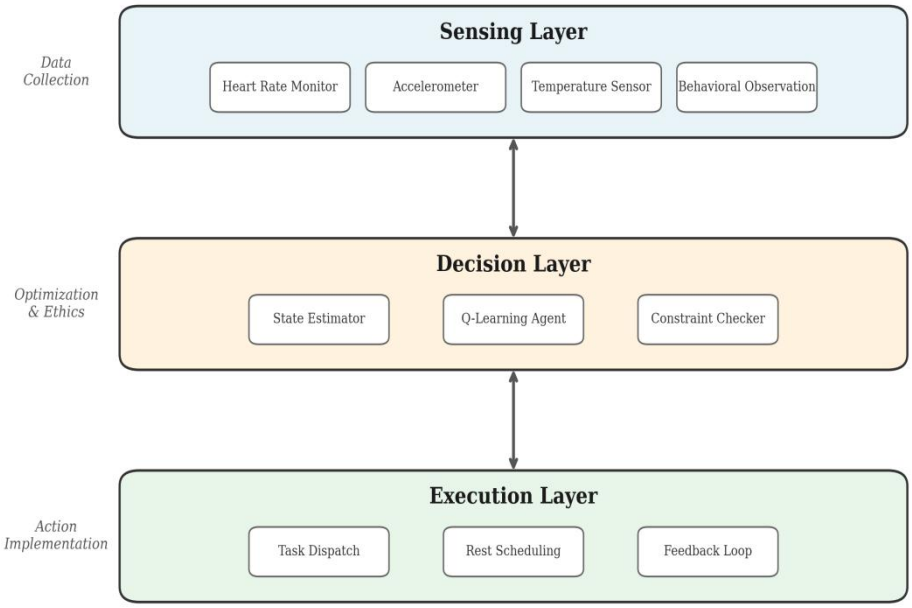

Figure 1. System Architecture

The Three Functional Layers of the System are: Sensing Layer: Acquire physiological and behavioural data from wearable devices, such as heart-rate monitors, accelerometers, and temperature sensors, as well as behaviour observations by handlers. Raw sensor data streams into the decision layer, and then the state estimator computes normalised stress indicators, fatigue indices and recovery deficits for each dog. The Q-learning agent obtains the current state estimate and outstanding task information, then generates an allocation decision that meets the constraints specified by the constraint checker for ethics and skills. The execution layer will perform the relevant operations according to a dispatch protocol and schedule the deployment and mandatory rest time of the task. Feedback loops return observed results to update state-estimate and improve value-function-approximation.

3.3. State Space Design

The state vector for each dog *i* at time t is defined as:

$$s_{i,t} = [sigma_{i,t}, phi_{i,t}, rho_{i,t}, kappa_{i,t}, tau_{i,t}] \tag{3}$$

where $sigma_{i,t}$ represents the normalized stress indicator, $phi_{i,t}$ represents the normalized fatigue level, $rho_{i,t}$ represents the recovery deficit accumulated over recent work periods, $kappa_{i,t}$ represents the skill proficiency vector for different task categories, and $tau_{i,t}$ represents the time since last rest period. All components are normalized to [0,1] intervals to facilitate learning stability.

The stress index sigma is derived from a group of physiological data.

$$sigma_{i,t} = w_1 c_{i,t} + w_2 h_{i,t} + w_3 b_{i,t} \tag{4}$$

where $c_{i,t}$ represents normalized cortisol level, $h_{i,t}$ represents normalized heart rate elevation, $b_{i,t}$ represents normalized behavioral stress score, and w_1, w_2, w_3 represent positive weights summing to unity. This composite indicator provides a scalar measure of overall stress burden that drives both the reward function and ethical constraints.

Table 2 shows the four types of tasks in the simulation. Patrol tasks have a moderate intensity and a short duration. Detection tasks have a relatively low physical intensity but are cognitively demanding and prolonged. Apprehension tasks have the highest stress impact because of the risk of physical confrontation. Search tasks have different durations in different areas.

Table 2. Task Type Characteristics

Task Type	Intensity Index	Mean Duration Hours	Cognitive Demand	Physical Demand
Patrol	0.35	4.0	Moderate	Moderate

Detection	0.25	6.5	High	Low
Apprehension	0.75	2.0	High	Very High
Search	0.45	5.0	Moderate	High

3.4. Action Space and Feasibility Constraints

The set of all feasible assignment matrices under the operating constraints constitutes the action space. Feasibility requires that each task is assigned to exactly one qualified dog, each dog receives at most one assignment per time step, and dogs exceeding the stress threshold are automatically assigned to rest. The feasible action set is formally given by:

$$\begin{aligned} \sum_{i=1}^{N} a_{i,j,t} &= 1, && j \in T_t \\ \sum_{j=0}^{M} a_{i,j,t} &\le 1, && i = 1,\ldots,N \\ a_{i,j,t} &= 0, && \sigma_{i,t} > \sigma_{\max}, j = 1,\ldots,M \\ a_{i,j,t} &= 0, && \kappa_{i,t}^{\tau(j)} < \kappa_{\min}^{\tau(j)} \end{aligned} \tag{5}$$

where N is the total number of dogs, M is the total number of operational tasks, $a_{i,j,t}$ is the assignment indicator equal to 1 if dog i is assigned to task j at time t and 0 otherwise, T_t is the set of active tasks at time t, $\sigma_{i,t}$ is the stress indicator of dog i at time t, σmax is the maximum permissible stress threshold, $\kappa_{i,t}^{\tau(j)}$ is the skill proficiency of dog i at time t for the task type $\tau(j)$ corresponding to task j, and $\kappa_{\min}^{\tau(j)}$ is the minimum skill requirement for that task type. First of all, ensure that each active task is linked to only one dog. The second is to prevent multiple allocations. The third equation indicates that dogs must rest when the stress exceeds a certain limit. The fourth is to have suitable skills for all the work.

3.5. Reward Function Design

A multi-objective form is used to combine operational efficiency and welfare costs in the reward function. The immediate reward of the state-action pair *(s, a)* is:

$$R(s,a) = alphaR_eff(s,a) + betaR_welfare(s,a) + deltaR_fair(s,a) \tag{6}$$

where R_eff represents the operational effectiveness component, $R_welfare$ represents the animal welfare component, R_fair represents the workload equity component, and alpha, beta, delta represent scalar weights satisfying alpha + beta + delta = 1.

The operational effectiveness component measures task completion quality:

$$R_eff(s,a) = sum_{i=1}^{N} sum_{j=1}^{M} a_{i,j,t} q_{i,j} eta_j \tag{7}$$

where $q_{i,j}$ represents the predicted success probability of dog i on task j based on skill match and current condition, and eta_j represents the operational priority weight of task j.

The welfare component penalizes stress elevation and rewards recovery:

$$R_welfare(s,a) = sum_{i=1}^{N} [lambda(sigma_{i,t} - sigma_{i,t+1}) - mu \max(0, sigma_{i,t+1} - sigma_t \arg et)] \tag{8}$$

Lambda is the reward coefficient for stress reduction; mu is the penalty coefficient for stress exceeding the target level; and $sigma_t \arg et$ is the desired stress maintenance level. This design will actively promote choices that reduce stress and discourage excessive increases in stress.

The fairness component discourages chronic overutilization of specific dogs:

$$R_fair(s,a) = -nusum_{i=1}^{N} (tau_{i,t} - tau_bar_t)^2 \tag{9}$$

where nu represents the fairness penalty coefficient, $tau_{i,t}$ represents the cumulative work time for dog i, and tau_bar_t represents the mean cumulative work time across all dogs. This quadratic penalty

drives equitable workload distribution.

Table 3 presents the hyperparameter settings and simulation parameters. The ethical stress threshold $sigma_max$ is set at 0.75, beyond which mandatory rest is enforced. The target stress level $sigma_target$ is 0.35, representing the desired equilibrium. Weight parameters alpha, beta, and delta are set to 0.5, 0.35, and 0.15 respectively, prioritizing operational effectiveness while maintaining substantial welfare influence

Table 3. Model Hyperparameters and Simulation Settings

Parameter	Symbol	Value
Learning rate	*alpha_lr*	0.10
Discount factor	*gamma*	0.95
Initial exploration rate	*epsilon_0*	0.30
Minimum exploration rate	*epsilon_min*	0.05
Exploration decay constant	*kappa_e*	0.001
Ethical stress threshold	*sigma_max*	0.75
Target stress level	*sigma_target*	0.35
Effectiveness weight	*alpha*	0.50
Welfare weight	*beta*	0.35
Fairness weight	*delta*	0.15
Number of dogs	*N*	8
Simulation horizon	*T*	30
Mean daily task arrivals	*lambda_task*	3.5

3.6. Ethical Intervention Mechanisms

Ethical interventions operate at the two levels of the model. Hard constraints are used to prevent the assignment of harmful values at the constraint level. Soft incentives at the level of reward shaping guide the policy towards welfare-promoting allocations. The set of ethical constraints *C* is as follows:

$$C = ainA_feasible : sigma_{i,t+1} <= sigma_maxforalli, andtau_{i,t} <= tau_maxforalli \tag{10}$$

where $sigma_max$ represents the absolute stress ceiling, and tau_max represents the maximum continuous work duration. These constraints implement deontological prohibitions against dangerous overwork.

Reward shaping adds a long-term welfare-foresight component to the base reward. A Shaped Reward is:

$$R_shape(s,a,s') = R(s,a) + gammaphi(s') - phi(s) \tag{11}$$

where *phi* represents a potential function estimating the long term welfare value of states. The potential function is defined as:

$$phi(s) = -zetasum_{i=1}^{N} sigma_{i,t}^2 \tag{12}$$

where zeta is a positive scaling constant. The potential function assigns a higher value to states with a lower total stress and thus motivates the policy to select a path that keeps the workforce in a low-stress environment. The shaped reward meets the condition of policy invariance, and thus the optimal policy with R_shape is the same as that obtained using *R*.

3.7. Q Learning Algorithm

The model employs a tabular Q learning algorithm to estimate the optimal action value function. The Q value update rule is:

$$Q(s,a)leftarrowQ(s,a) + alpha_lr[r + gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{13}$$

where *Q(s, a)* represents the estimated value of taking action *a* in state *s*, *alpha_lr* represents the learning rate controlling update magnitude, r represents the immediate reward received, gamma represents the discount factor, s' represents the subsequent state, and $\max_{a'} Q(s',a')$ represents the maximum estimated value of the next state. This update propagates reward information backward through the state space, enabling the agent to learn long term consequences of allocation decisions.

Due to the large state space, binning of the continuous state variables is used to perform state aggregation. The five levels of the discretised stress indicator sigma are: very low, low, moderate, high and critical. Discretize the three levels of fatigue phi. The three levels of recovery deficit rho are discretised. This aggregation results in a small state space and keeps decision-relevant distinctions.

Exploration is carried out by epsilon-greedy action selection. A random feasible action is selected with probability ε. With probability $1-\varepsilon$, the action that maximises the current Q estimate is selected. The exploration rate decreases with the training episodes according to:

$$epsilon_t = epsilon_min + (epsilon_0 - epsilon_min)\exp(-kappa_e t) \tag{14}$$

where $epsilon_t$ represents the exploration rate at episode t, $epsilon_min$ represents the minimum exploration rate, $epsilon_0$ represents the initial exploration rate, and $kappa_e$ represents the decay constant. This schedule ensures adequate early exploration while converging to exploitation as value estimates improve.

Figure 2 is the training convergence of the Q-learning agent. The x-axis is the training episodes from 0 to 5000. The left is the smoothed episode reward. The curve starts around -50 and then rises; thus, stress violations and poor task matching in random exploration are severely penalized. The reward increases monotonically in the first 2000 episodes as the agent learns a feasible policy. After 3500 episodes, the curve has reached a plateau near +85 and is close to convergence of the policy. The shaded area shows one standard deviation over 20 independent training runs and is stable in terms of learning behaviour.

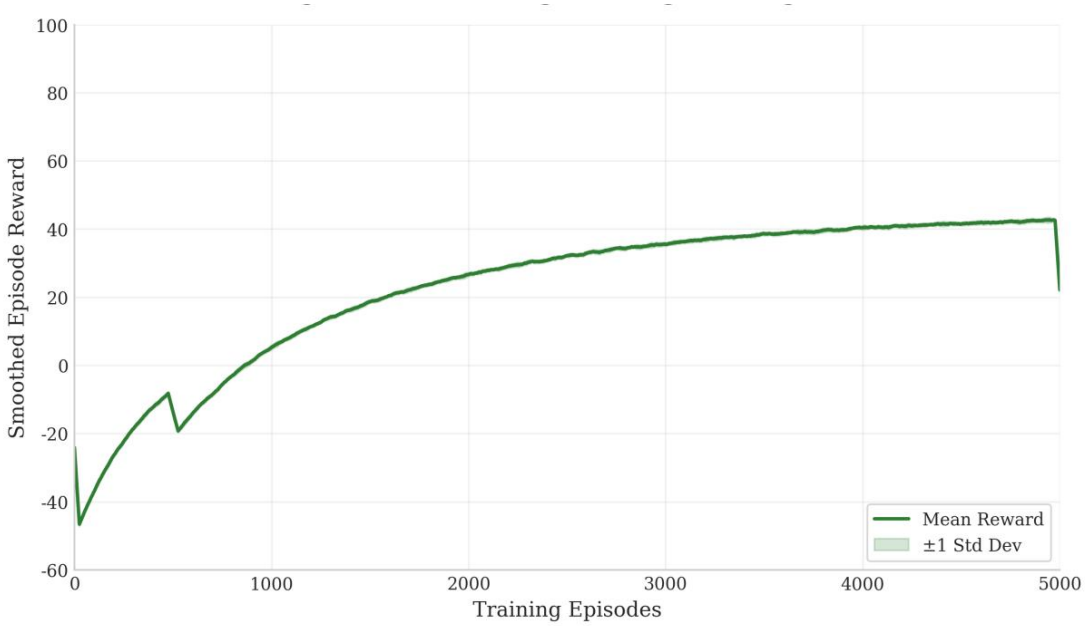

Figure 2. Q Learning Training Convergence

Figure 2 is the learning curve after 5000 training iterations. The episodes on the horizontal axis are from 0 to 5000. The left vertical axis is the smoothed episode reward, and it ranges from -60 to 100. Near -50, the curve is relatively low; it rises sharply around episode 1000, increases gradually up to episode 3500, and then flattens out near +85. A shaded envelope around the curve indicates plus or minus one standard deviation of the results of 20 independent runs. The convergence pattern shows that the agent has learned to balance operational efficiency and welfare costs reasonably well in the training period.

3.8. Simulation Environment

Build a discrete-event simulation environment to test the new model. The environment is a police canine unit with N=8 dogs and operates for 30 days, taking one day per step. Task arrivals follow a Poisson process with a mean rate of lambda_task = 3.5 per day. Task types are patrol, detection, apprehension and search, and they each have different intensity and duration distributions.

Individual dog parameters are sampled from distributions calibrated by empirical data of police canine units. The stress sensitivity theta_i is drawn from a Beta distribution with parameters 2 and 5,

and individual differences in stress response to the same task are thus generated. Recovery rate omega_i is uniformly distributed in the range [0.15, 0.35], and it represents the proportion of stress recovery after a rest period. Skill proficiency vectors are initialized by training records and change with experience.

Round-robin scheduling is a baseline comparison method that cycles through dogs in a fixed order; random assignment distributes tasks uniformly; and a greedy heuristic assigns each task to the dog with the highest predicted success probability without considering future stress.

3.9. Experimental Results

Train the Q-learning agent for 5000 episodes with a learning rate of alpha_lr = 0.1, a discount factor of gamma = 0.95, an initial exploration rate of epsilon_0 = 0.3, and a decay constant of kappa_e = 0.001. The moving average of episode rewards was used to judge training convergence.

Figure 3 shows the cumulative stress paths for the allocation methods over a 30-day period. The left y-axis is the mean stress index of the dog unit. Round-robin shows a gradual rise in stress with small daily fluctuations and reaches 0.68 by day 30. The greedy heuristic has higher volatility and reaches 0.71 due to the excessive use of high-skill dogs. Random Assignment shows a relatively small increase to 0.64. The proposed reinforcement learning method keeps the stress between 0.32 and 0.45 over the entire horizon and shows adaptive rest periods as periodic drops.

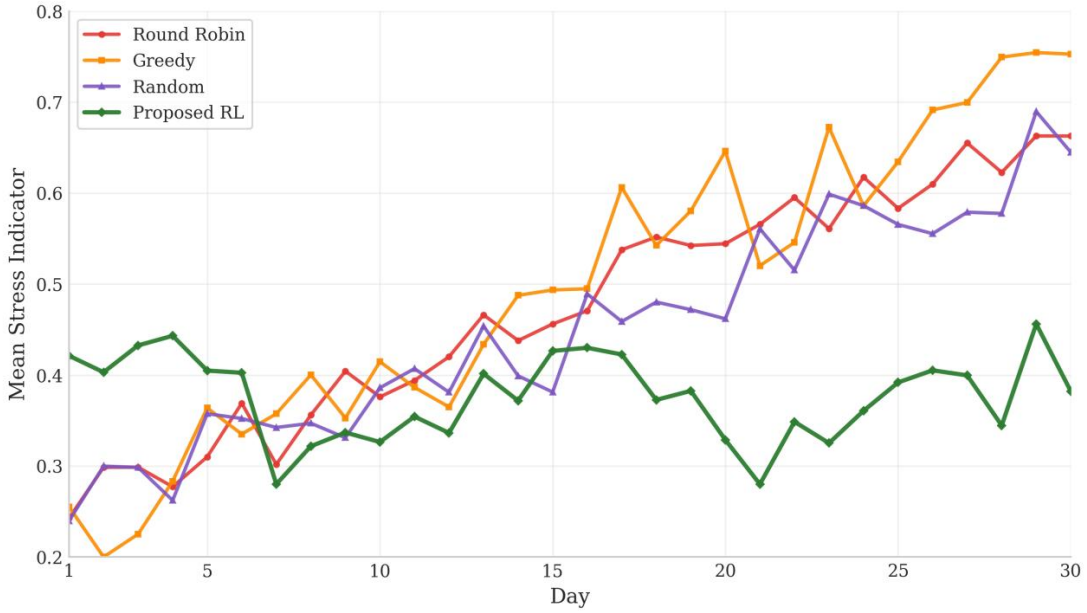

Figure 3. Cumulative Stress Trajectories Based on Allocation Method

Figure 1 shows the average unit stress of the four allocation methods over 30 days. Days 1-30 are displayed on the horizontal axis. The vertical axis is the mean stress index (0.2 - 0.8). The first round-robin path is 0.25, and then with small increments, it gradually increases to 0.68. The first point of the greedy trajectory is 0.25; it has a relatively large daily fluctuation and reaches a high of 0.71. Random Trajectory increases to about 0.64 and shows irregular fluctuations. The range of the reinforcement learning trajectory is [0.32, 0.45], and an adaptive downward spike occurs at the scheduled recovery time. The disparity between the new way and the old one is relatively small in the first 10 days.

Figure 4 shows the task completion rate and welfare indicators of the methods. Panel A is the proportion of tasks completed successfully. Reinforcement learning achieved a completion rate of 94.3%, which was higher than those for greedy (89.1%), round robin (87.6%) and random assignment (82.4%). Panel B is the mean number of mandatory rest interventions triggered per dog. Reinforcement Learning needs fewer regular interventions than the greedy and round-robin methods because the policy can avoid exceeding the stress limit proactively. Panel C shows the coefficient of variation in work hours, which indicates inequity. The lowest value of the reinforcement learning method is 0.18, and it is relatively balanced compared with 0.31 for greedy.

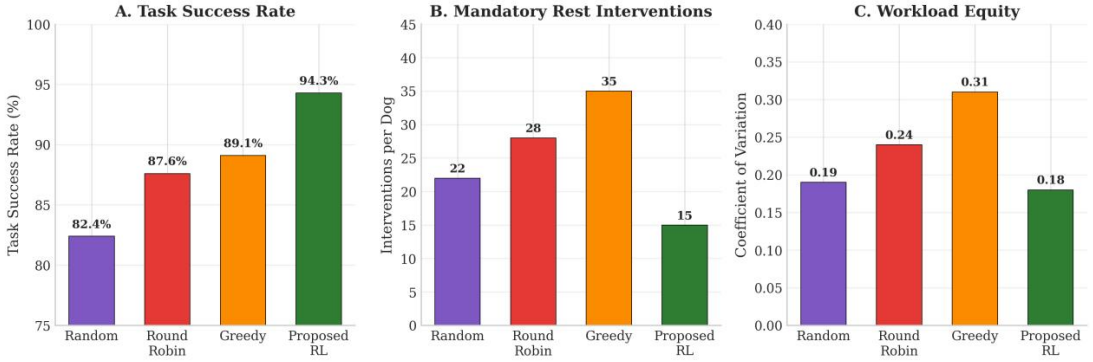

Figure 4. Operational and Welfare Metrics Comparison

The three panels in the figure are the methods. Panel A shows the task completion rates in a bar chart. Random assignment is 82.4 per cent, round robin is 87.6 per cent, greedy is 89.1 per cent, and reinforcement learning is 94.3 per cent. Panel B shows the required rest intervention per dog as vertical bars. Random needs 22, Round Robin needs 28, Greedy needs 35, and Reinforcement Learning needs 15. Panel C is the workload variation coefficient as a bar chart. Random shows 0.19, Round Robin shows 0.24, Greedy shows 0.31, and Reinforcement Learning shows 0.18. Reinforcement Learning has achieved a high completion rate, required minimal intervention, and evenly distributed the work.

Table 4 is the general performance comparison. Reinforcement learning has a relatively high task completion rate of 94.3% and a low mean stress of 0.38. It also has the lowest number of stress threshold violations at 12 instances across all dogs and 30 days, compared with 47 for round robin and 61 for greedy. Under the new way, the mean recovery time between demanding tasks is 2.1 days; under round robin, it is 1.4 days, so the new way can achieve more thorough physiological recovery.

Table 4. Performance Comparison Across Allocation Methods

Metric	Round Robin	Random	Greedy	Proposed RL
Task completion rate percent	87.6	82.4	89.1	94.3
Mean stress indicator	0.61	0.58	0.64	0.38
Stress threshold violations	47	38	61	12
Average recovery days	1.4	1.6	1.2	2.1
Workload variation coefficient	0.24	0.19	0.31	0.18
Mandatory rest interventions	28	22	35	15

Sensitivity Analysis of the Effect of Ethical Weight Parameter Beta on Outcomes. As beta increases from 0.1 to 0.5, mean stress decreases from 0.52 to 0.31, and task completion decreases from 96.1 per cent to 91.8 per cent. The parameters chosen by the administrator from the trade-off curve will be in line with their institution's ethical standards. At beta = 0.35, the model is nearly Pareto optimal in terms of welfare and efficiency.

An increase in the size of the uncertainty region. When the estimated stress-sensitivity parameters have a 20% error, the reinforcement learning method performs reasonably well and keeps a 92.7% completion rate and a mean stress of 0.42. Robustness arises from the exploratory nature of Q-learning; thus, it adapts to the observed environment rather than assuming it.

4. Summary

This paper introduces and verifies a reinforcement learning-based model for dynamic police dog task allocation that explicitly includes stress reduction and ethical intervention as the main optimisation goals. The system of mathematics presented here considers each working dog an independent agent, and the physiological condition of these agents is influenced by task assignment and rest. By incorporating the costs of animal welfare directly into the reward function and setting a hard constraint on stress exposure, the allocation decision of the model will satisfy both operational needs and ethical requirements for sentient beings. Based on the experimental results, the first way to reduce stress is better than the old schedule, and at the same time, through intelligent allocation of tasks, the dogs' long-term working ability has also been preserved.

The application of this research will also benefit other places where animals are used for work under human direction, such as search and rescue organisations, military working-dog programs,

assistance-dog deployment, etc. Methodologically, it has been shown that reinforcement learning can be applied to scenarios where the agents being scheduled are not the learners, but rather the subjects of welfare protection by the learning system. Inversion of the typical reinforcement learning paradigm, in which the learning agent acts in its own interest, raises serious problems of agency, consent and representation for animals in algorithmic systems.

Several deficiencies have been identified and the future direction of research proposed. The current model is a discrete-state model that cannot account for variations in physiology relevant to the clinic. Deep reinforcement learning architectures that can handle continuous state representations may be more suitable for individual differences but require significantly larger training datasets. Although the simulation environment is based on real data, it does not fully represent the complexity of actual police work, and tasks and environments in practice are often vague and varied. Field validation with real dog units can improve the application value of the above results. In addition, the ethical system that will be applied here will focus mainly on stress reduction and workload fairness, and other types of welfare, such as behavioural enrichment, socialisation and environmental comfort, have not yet been included.

Future research could explore multi-objective optimisation formulations that represent welfare as a vector rather than a single value, thus showing the trade-off among various quality-of-life indicators explicitly. Transfer learning can be used to have a model trained on a group of dogs quickly adapt to a new group with different characteristics. Prediction of health problems or injuries according to changes in body conditions may be added to strengthen the protective function of the allocation system. Finally, participatory Design can be employed to involve handlers, veterinarians and animal behaviourists in setting the specifications for reward functions and constraints to enhance the legitimacy and acceptance of algorithmic management in this sensitive area. Looking ahead, the three parties will continue to work together: computer scientists, animal welfare organisations and law enforcement, etc., to ensure that new technologies are used responsibly to protect both the public and the animals in their care.

References

1. Birnbaum, A., et al. (2019). Multi agent reinforcement learning for distributed task allocation. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 34-41.
2. Cobb, M. L., et al. (2020). Physiological stress in detection dogs during operational deployments. Applied Animal Behaviour Science, 230, 105001.
3. Darby, W., et al. (2019). Stress and recovery in police dogs during tactical operations. Journal of Applied Animal Welfare Science, 22(4), 345-358.
4. Haverbeke, A., et al. (2008). Training methods of police dog handlers and their effects on the behavior and cortisol responses of police dogs. Journal of Veterinary Behavior, 3(2), 75-82.
5. Mancini, C. (2011). Animal computer interaction: A manifesto. Interactions, 18(4), 69-73.
6. Mao, Y., et al. (2022). Deep reinforcement learning for dynamic resource allocation in emergency response systems. IEEE Transactions on Intelligent Transportation Systems, 23(5), 4123-4135.
7. Mnih, V., et al. (2015). Human level control through deep reinforcement learning. Nature, 518(7540), 529-533.
8. Rault, J. L., et al. (2020). Physiological biomarkers of animal welfare: A multi species perspective. Frontiers in Veterinary Science, 7, 589.
9. Rooney, N. J., & Cowan, S. (2011). Training methods and owner dog interactions: Links with dog behaviour and welfare. Applied Animal Behaviour Science, 132(3-4), 169-177.
10. Rooney, N. J., et al. (2017). The welfare of brachycephalic breeds in police work. Animal Welfare, 26(2), 215-223.
11. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT Press.
12. Watkins, C. J., & Dayan, P. (1992). Q learning. Machine Learning, 8(3), 279-292.
13. Webster, J. (2016). Animal welfare: Freedoms, dominions and a life worth living. Animals, 6(6), 35.
14. Zhang, L., et al. (2021). Reward shaping in multi objective reinforcement learning for sustainable resource management. Neural Computing and Applications, 33(8), 3789-3801.