Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

A Multi-Intelligent Body Simulation Study of Minor Crime Prevention

Jin Mei[1,✉], Yichen Zhou[2], Shanxin Zhang[1]

[1] *Jiangnan University, Wuxi, Jiangsu, 214122, China*
[2] *Wuxi Taihu University, Wuxi, Jiangsu, 214064, China*

ABSTRACT
This paper constructs a multi-agent simulation model to study and prevent juvenile delinquency. A multi-agent reinforcement learning model is constructed according to reinforcement learning theory to simulate the behavioral decision-making process of minors in different social environments. By introducing the NashQ algorithm, it simulates the minors' strategic choices when facing the temptation of crime. In the simulation experiments, the NashQ algorithm meets the convergence requirements of the model, and only 1/3 of the training times are needed to achieve the stability of the simulated environment. Among them, family factors, school factors and social factors all affect the stability of the prevention effect. Good family environment, high quality teaching conditions and healthy social atmosphere can effectively prevent juvenile delinquency.

Keywords: Multi-intelligence, Multi-agent reinforcement learning, NashQ algorithm, Crime prevention

1. Introduction

Minors are citizens under the age of eighteen, they are the future of the country, the hope of the nation and the dream of the family, and they are related to the great rejuvenation of the Chinese nation and the happiness and peace of hundreds of millions of families. Minors are characterized by the fact that their minds are not yet well developed, and their outlook on life and values are still in the state of gradual enrichment, not mature enough [1-2]. Under the background of globalized society of information development, minors are very easy to be influenced, such as very easy to be tempted by some undesirable products produced by social development, and many kinds of bad customs produced by diversification of ideas also have serious negative impact on the physical and mental development of minors [3-6]. When all kinds of problems come together, it intensifies a social problem that can not be ignored, that is, juvenile delinquency.

As minors are in the stage of exploration and growth, plasticity is stronger, starting from the

✉ Corresponding author.
E-mail address: taihu1984@163.com (J. Mei).
Received 25 February 2024; Revised 10 May 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-504

prevention aspect is obviously more effective than punishing after committing a crime [7-9]. The so-called crime prevention is a kind of activity that aims at the specific causes of the crime phenomenon, takes appropriate countermeasures to exclude the relevant factors that cause the crime, and corrects and prevents those who are likely to commit the crime, and then can effectively reduce or completely eliminate the crime [10-13]. Multi-intelligent system (MAS), as an important means of complex system analysis and simulation, provides a brand new idea for research related to juvenile crime prevention [14-16]. Compared with the general method of simulation research, the MAS method pays more attention to the simulation of human behavioral decision-making, "bottom-up" to explore the role of the dynamic micro subjects on the system's macroscopic pattern and function, so it can be more pertinent to explain and predict the problems occurring in the complex juvenile delinquency prevention [17-20].

The study adopts a multi-agent reinforcement learning framework, in which each agent represents a minor with attributes such as age, gender and family background, and learns and makes decisions based on individual experience and social feedback in a simulated environment. The NashQ algorithm is chosen as the reinforcement learning algorithm in a multi-intelligent agent environment, and the crime prevention strategy is defined as a non-cooperative multi-agent system, and its investment and punishment models are given. Crime prevention strategies are investigated using the Nash equilibrium based Q learning algorithm. Experimental simulation of the crime prevention strategy using the NashQ algorithm is conducted to investigate the stability of the prevention effect under different causes of juvenile delinquency.

2. Causes of juvenile delinquency

2.1. Family factors

Parents are the first teachers of their children, and the various behaviors and words shown by parents have an impact on the generation of values of children at all times. The younger a child is, the more dependent he or she is on his or her parents, and the more profound the influence is on the child. Family disharmony will bring indelible psychological shadows to minors, leading to the child's three wrong views. In some families, the parents often quarrel, bringing harm to the children's hearts. Some parents have a violent way of education, put too high requirements for their children, once the children make mistakes or poor academic performance, they will be severely punished, minors in such a force of deterrence, often appear rebellious. Even some parents themselves have many bad habits, resulting in children also developed bad habits. Some parents overly pampered children, excessive pursuit of enjoyment of the material life, neglected to cultivate the spiritual world, long-term, the child becomes a gold-digging. Some parents don't really take their legal responsibility seriously, they are so immersed in their own working environment that they don't have enough time to discipline their children. In the long run, their children can't get the feeling that they are valued at all, and they will develop bad habits and go astray.

2.2. School factors

1) Lack of educational content. In the process of school education and teaching, it violates the laws of education, and purely carries out exam-oriented education, neglecting the comprehensive cultivation of students, which brings about an impact on the healthy growth of minors and leads to the problem of crime. Firstly, ideological and moral education and legal education are slighted. Secondly, the psychological education is lightly regarded, the psychological problems of minors are very serious, most of the schools did not pay attention to the psychological education of minors, and the schools seldom offer special psychological education courses, which leads to the psychological problems of the students can not be solved in a good way [21]. If it can not be solved in time, it may lead to students using extreme ways to vent out, resulting in crime. Finally, sex education is lightly regarded, with the

physiological development, minors gradually begin to mature, most schools also did not provide students with correct sex education, resulting in students do not understand sexual knowledge, curious about pornographic websites, by the influence of bad education, thus increasing the chance of crime.

2) Deficiencies in the teaching management system. First of all, there is a misunderstanding of the guiding ideology of teaching, most schools are only concerned about the rate of advancement, do not pay attention to the all-round development of students, resulting in students have no interest in learning. Secondly, the school management system is not perfect, the school can't fully understand the students' trend, the students are influenced by the bad ideas outside the school, and they have no knowledge of the illegal behavior. Finally, the communication mechanism between schools and parents is not sound, especially in rural areas, where there is no perfect visiting system and little communication between schools and parents, so that they are unable to understand the social relations and family situation of students in a timely manner, and are unable to stop students from violating the law in a timely manner.

2.3. Social factors

1) Unhealthy social culture and spiritual products. Cyberspace is filled with a lot of information that is harmful to the health of young people's minds, and some Internet cafes have lax control and chaotic environments, laying a hidden danger for juvenile delinquency. Nowadays, the bad network culture causes the number of juvenile delinquency to increase gradually, but there is no effective management, all kinds of pornographic vulgar content often appear, to the weak discernment ability of the minors easy to produce the thought of misleading, most of the minors will be subjected to the bad influence of these websites.

2) Misleading by bad social values. Various phenomena of gold worship, corruption and embezzlement arise, and the media report that the remuneration of stars gradually climbs, which brings great influence for minors and directly affects their thinking, leading to their great material desires without sufficient material foundation, blindly pursuing an extravagant life, and attaching importance to the material life that is not in line with the actual situation, which leads to minors' misunderstanding of the values, and results in their Because of the desires that do not correspond to the actual situation to produce robbery and other illegal and criminal behavior.

3. Multi-intelligence body simulation modeling for juvenile delinquency prevention

3.1. Multi-intelligent body reinforcement learning

3.1.1. Enhanced learning. Reinforcement learning is an online learning technique that is very distinct from supervised and unsupervised learning. The basic model is shown in Figure 1. "Exploration-utilization" is regarded as a learning process, in which the learning system first perceives the state of the environment, takes a certain action to act on the environment, and the state changes after the environment accepts the action, and at the same time gives a return value (reward or punishment), which is fed back to the multi-agent system, and the multi-agent system selects the next action according to the reinforcement signal and the current state of the environment, and the principle of selection is to make the return value increase continuously [22].

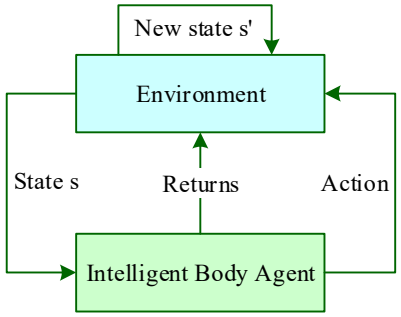


Fig. 1. Reinforcement learning frame structure

The interaction process at each moment when the Intelligent Body Agent interacts with the environment is as follows:

1) The Intelligent Body Agent acquires the state S_t at a moment t.

2) Based on the current state and accumulated returns $R(t)$, the Intelligent Body Agent selects action a to interact with the environment.

3) When the action a selected by the intelligent agent is executed, the environment changes. The environment shifts from the current state s_t to the next new state s_{t+1}. Giving an immediate reward $R(t)$, also known as the reward value.

4) Immediate return $R(t)$ feedback to the intelligent body Agent, $t \leftarrow t+1$.

5) Turn to step 2 and stop the loop if the new state is the end state.

Where the immediate return $R(t)$, is determined by the state of the environment s_t and the output of the intelligent body Agent a_t. $a \in A$, A for a set of movesets.

After the Agent performs an action, the tendency of the intelligence to produce this action again is strengthened when the signal from the environment feedback is positive. Conversely Agent's tendency to produce this action weakens. Therefore the condition for an Agent to make behavioral choices is that the feedback value generated by the environment after executing an action is positive, and the Agent should be able to continuously strengthen this signal through repeated trials on the system [23]. The learning objective of the intelligent Agent is to maximize the cumulative payoff value. The evaluation function $V(t)$ is a measure of the long-term return and consists of three main return expressions:

1) Finite range model: $V(t)$ is the sum of cumulative returns of the Intelligent Body Agent in a finite number of states. t is the sampling moment, n is the total number of states the intelligent body runs from t the moment to the end, and n can be not predetermined:

$$V(t) = r_t + r_{t+1} + \ldots\ldots + r_{t+n-1} \tag{1}$$

2) Discounted return (infinite state model): it is the sum of cumulative returns to the intelligent Agent within an infinite number of states:

$$V(t) = \sum_{a=0}^{\infty} \gamma^n r_{t+n} \tag{2}$$

γ is the discount factor, usually $0 \le \gamma < 1$. By regulating γ, it is possible to control the extent to which the multi-intelligent body learning system considers the short- and long-term outcomes of its own actions.

3) Average return model: reinforcement learning uses a third criterion, the average of the cumulative returns of the intelligent Agent, standardized as:

$$V(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n} r_{t+k} \tag{3}$$

Of the three evaluation expressions above, the one with the highest chance of being used is the discounted return cumulative function.

The problem of interaction between an intelligent agent and the environment can be modeled as a Markov decision model. First consider a finite stochastic process with the state of the environment denoted as $S=\{s_0,s_1,\ldots,s_n\}$. Denote by a random variable $a(t)$ the action taken by the learning system at time step t. Denote by $P_{s_i}(a_t)$ the transfer probability of the learning system to move the environment from state s_i to state s_j at time step t due to the probability distribution of taking action a_t, where $a_t \in A$, $s_t \in S$. Then the transfer probability is:

$$P(a_t)=p(s_{t+1}=s_j \mid s_t, a_t=a) \tag{4}$$

The multi-agent learning system should satisfy:

1) The selection of states should reflect the Markov property, i.e., the state at the next moment is only related to the current state, but not to the previous historical state. That is, there is no posteriority.

2) The selection of actions should both make all states traversable and ensure that there is a definite state transfer law, i.e., traversability. At the same time, the goal to be accomplished by the intelligent body Agent is accomplished by a series of actions.

3) The payoff function should be chosen in such a way that the Intelligent Body Agent can find the optimal strategy π^*.

3.1.2. Multi-Agent Reinforcement Learning:

1) Basic model of multi-agent reinforcement learning. The basic model of multi-agent reinforcement learning is shown in Figure 2. The purpose of intelligent body Agent learning is the process of continuous interaction with the environment in order to accomplish a predetermined task. Sometimes it is necessary for multiple Intelligent Bodies to jointly collaborate to acquire together, but when the action of a single Intelligent Body Agent cannot accomplish the predetermined goal, the Intelligent Body Agent needs to consider adopting the corresponding learning strategy to complete the task that it should accomplish. The study simulates the behavioral decision-making process of minors in different social environments. Each Agent represents a minor with attributes such as age, gender and family background, and learns and makes decisions based on individual experience and social feedback in the simulated environment.

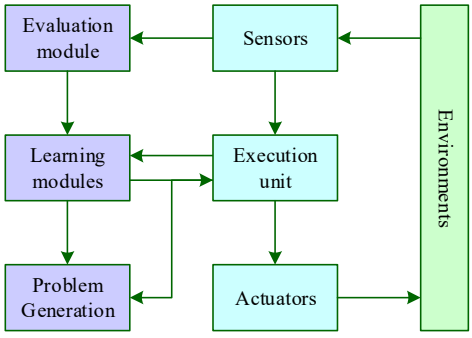


Fig. 2. Multi-agent Learning Mode1

2) Components of a Multi-Agent Learning System. Reinforcement learning system can be modeled as a mathematical model using Markov decision process or semi-Markov process. Reinforcement learning system in addition to the intelligent body and environmental factors, but also contains the following elements: strategy, reward function, value function, which is also the basic elements of artificial intelligence problems.

Strategypolicy, a strategy is a collection of decision sequences. When an intelligent body interacts with the environment, it needs a certain strategy to determine which action the intelligent body should choose in a certain state. The strategy for an intelligent body to interact with the environment may be a simple table or a simple mathematical function. And in some complex cases, the strategy may be some kind of intelligent algorithm for solving practical problems: e.g., greedy strategy.

Return function, refers to the immediate evaluation of the intelligent's behavior by the environment during the interaction between the intelligent and the environment. It is a constant that indicates, to some extent, how good or bad the Agent performs an action in a certain state. The payoff function is random in nature.

Value function, a quantitative indicator of the merit of a realized process. It is a quantitative function defined over the whole process and all subsequent sub-processes, usually denoted by $V(s)$. For an indicator function to constitute a dynamic environment, it should be separable and satisfy recursive relationships. There are three main forms of common value functions:

1) The indicator of a process and any of its sub-processes is the sum of the indicators of the stages it contains.

2) The metrics of the process and any of its sub-processes are the product of the metrics of the stages it contains.

The value function has different meanings in different problems, e.g., the value function in a multi-intelligence environment represents the expected value of the cumulative reward received by an intelligent from the time it starts to perform an action in state s_i and adopts a subsequent strategy in state π until it finally reaches the goal:

$$\begin{aligned} V(s) &= E\left\{r_{i+1} + \gamma r_{i+2} + \ldots \mid s_i = s\right\} \\ &= E\left\{r_{i+1} + \gamma V\left(s_{i+1}\right) \mid s_i = s\right\} \end{aligned} \tag{5}$$

For any strategy π, define the value function to be the expected value of the cumulative discounted reward at infinity, as described below:

$$V_{\pi}(s) = E_{\pi}\left(\sum_{i=0}^{\infty} \gamma^i r_t \mid s_0 = s\right) \tag{6}$$

where r is the immediate reward, s denotes the state, and the discount factor $\gamma(\gamma \in [0,1])$ is a fixed value indicating the relative ratio of delayed to immediate reward. In some cases it is more important to record the state-action than the $Q(s,a)$ value alone, and in this paper we refer to the state-action value as the Q value. $Q(s,a)$ under policy π denotes the expectation of the discounted reward and that obtained by performing action a while in state s. The value of Q is then continuously estimated by the intelligent body through long-term observation throughout its life cycle.

3.2. Minor crime prevention strategy based on NashQ algorithm

This paper proposes a research on juvenile delinquency prevention strategies based on the NashQ algorithm. The algorithm is used to better understand the complexity of juvenile delinquency through computer simulation of interactions between intelligences, so as to design an effective prevention strategy.

3.2.1. Correlation Models:

1) Algorithm design objectives. In this paper, the simulation game of juvenile delinquency is designed as follows: each participant in the game is defined as an Agent, then the simulation game of juvenile delinquency can be defined as a multi-agent system. The traditional experimental simulation experiment design is more just around how early to make the experimental participants in the punishment to achieve convergence earlier. For computer experiment simulation, the same for the

experiment earlier convergence is also an important design goal, but at the same time we require the simulation can be closer to the actual negotiation, more humane. To make the experiment a simulation platform for juvenile delinquency prevention negotiations, the NashQ algorithm was applied to the study of juvenile delinquency prevention strategies.

2) Investment model. Set in a game environment of N Agent, the training of the Agent is carried out in the unit of "round". In any round of training, as long as the investment of 10 times, whether or not to achieve the goal of training is over, while all Agents set the initial state, and then start a new round of training. The goal of the game is that the total investment of all Agents reaches $€N*20$, in which each Agent is absolutely aware of the goal, i.e., the Agent is first able to recognize how much its own goal is, and at the same time is able to memorize its own investment and the goal achievement situation in the previous game. At the same time, the game requires each Agent to invest in each round of investment and can only invest $€0$, $€2$ or $€4$, Agent can know the total amount of money after each investment, but its investment in the investment before and can not get the investment amount of other Agents, the investment amount of specific by the Agent itself according to its prediction of the investment amount of other Agents, the current environment and the goal to decide.

3) Penalty model. At the end of each round of the game, if the total amount invested can reach or exceed $€N*20$, then all Agents will get all the money left in their accounts. On the contrary, if the game does not reach the goal, as a punishment they will lose all the money in their accounts with a certain penalty probability. Due to the individual selfishness of Agents, all Agents try to get more money at the end of the game, which requires them to invest as little as possible each time, thus triggering a game between individual and collective interests. However, in order for the game to continue, all Agents reinitialize their accounts after each round, but are able to remember the results of all previous rounds. When the goal is not reached, the penalty is enforced using roulette (a single 6-point color), where the penalty probability is p and n is a random number of points.

Penalty Scenario:

$$Penalties\begin{cases} n < p*6, \ Penalties \\ No\ punishment \end{cases} \tag{7}$$

3.2.2. Principles of the NashQ algorithm:

1) Introduction to the algorithm. The NashQ algorithm extends multi-agent learning to non-cooperative general and stochastic games where the interests of Agents are not completely opposed. In this case, Agents can observe each other and obtain information about other Agents' previous actions and rewards, etc., and learn their own Q function Q_t^i while predictively modeling other Agents' Q function $Q_i^j\left(j \in \{1,\cdots,n\} And\ j \neq i\right)$. If the optimal policy under Nash equilibrium is set to be $\pi^1(s),\cdots,\pi^n(s)$ and all Q values in the same state form a policy to be $Q_t^1(s),\cdots,Q_t^n(s)$ then the value function is defined to be $Agent_i$ the reward for choosing Nash equilibrium in state s, i.e.:

$$V^i(s) = NashQ_t^i(s) = \pi^1(s)\cdots\pi^n(s)Q_t^i(s) \tag{8}$$

Thus, the value of Q in the NashQ algorithm is updated to:

$$Q_{t+1}^i(s,\vec{a}) = (1-\alpha)Q_t^i(s,\vec{a}) + \alpha\left[r_t^i + \beta \cdot NashQ_t^i\left(s_{t+1}\right)\right] \tag{9}$$

The Nash equilibrium $\pi^1(s),\cdots,\pi^n(s)$ under game situation $Q_t^1(s),\cdots,Q_t^n(s)$ is a vector, while there may be more than one Nash equilibrium solution for the same game, and the values of its Nash equilibrium solutions are different. Therefore, selecting different equilibria will lead to different updates of the value of Q.

2) Action prediction method. The action prediction method is to predict future actions in a specific state by observing and recording the past changes of Agent actions under rewards and punishments, i.e., building a mathematical model describing the change rule of actions based on the historical data of changes in the action state, as a way of predicting future actions in a specific state. The basic form of the action prediction function is set to $I(s,a)$, which represents the probability that other Agents will perform action a in state s, and in state s, the action prediction function satisfies condition $\sum_{a\in A} I\left(s,a_i\right)=1$. Based on the different methods of evaluating actions to give action predictions, we list and compare the following three different action prediction methods: the largest prediction, equal probability prediction, and maximum prediction based on the hypothesis test of the law of distribution.

(1) Maximum prediction. The predicted probability of any state for each action is the same at moment t_0, and if n is the number of actions, the predicted probability is $1/n$. After that, the value of the predicted probability is updated according to Eq. (9) every time the other Agents perform an action, and the corresponding action with the largest predicted probability value is computed at the time of prediction, which is referred to as maximal prediction:

$$I_{t+1}^i\left(\vec{s_t},a_k^i\right)=\begin{cases} I_t^i\left(\vec{s_t},a_k^i\right)+\beta\sum\limits_{d_k\in A-\{d_k\}} I_t^i\left(\vec{s_t},a_t^i\right), If\ a_k^i=a_t^i \\ \left(1-\beta\right)I_t^i\left(\vec{s_t},a_k^i\right),\ Other \end{cases} \tag{10}$$

where a_k^i denotes the kth action in the $Agent_i$ action set A^i and β is the learning rate of the predictive learning model.

The number of probabilistic prediction functions is updated according to Eq. (9), which always makes the sum of the predicted probabilities for all a_k^i equal regardless of any moment, i.e:

$$\begin{aligned} \sum_{a_i\in A} I_{t+1}^i\left(\vec{s}_t,a_t^i\right) &= I_{t+1}^i\left(\vec{s_t},a_k^i\right)+\sum_{\alpha_k\in A-\{\alpha_k\}} I_{t+1}^i\left(\vec{s_t},a_t^i\right) \\ &= \left(I_t^i\left(\vec{s_t},a_t^i\right)+\beta\sum_{a_k^i\in A-\{\alpha_t^i\}} I_t^i\left(\vec{s_t},a_k^i\right)\right) \\ &+\sum_{a_k^d\in A-\{\alpha_i^d\}}(1-\beta)I_t^i\left(\vec{s_t},a_k^i\right) \\ &= I_t^i\left(\vec{s_t},a_t^i\right)+\beta\sum_{a_k^i\in A-\{\alpha_i^i\}} I_t^i\left(\vec{s_t},a_k^i\right) \\ &+(1-\beta)\sum_{a_k^i\in A-\{\alpha_i^i\}} I_t^i\left(\vec{s_t},a_k^i\right) \\ &= I_t^i\left(\vec{s}_t,a_t^i\right)+\sum_{\alpha_k^d\in A-\{\alpha^d\}} I_t^i\left(\vec{s}_t,a_k^i\right) \\ &= \sum_{a_k^i\in A} I_t^i\vec{s_t}a_k^i \end{aligned} \tag{11}$$

(2) Equal probability prediction. For equal probability prediction, when an Agent predicts what kind of action other Agents will take, it considers the probability that each action in the action set will be executed is equal. It does not need to observe and analyze the action selection and execution of other Agents, and does not consider the historical execution data of other Agents at all, thus greatly reducing the storage space requirement. Here the prediction of actions can be realized using the roulette method. If we set n as the size of the action set, the predicted probability of each action a being executed is:

$$P(a) = 1/n \tag{12}$$

(3) Maximum prediction based on distribution law hypothesis testing. As for the maximum prediction based on distribution law hypothesis testing, it is first assumed that, at the initial moment t_0, the value of the predicted probability of any action is the same for every state, i.e., $1/n$.

The statistical law of distribution of variable x is known, and if it needs to be de-fitted, then whatever theoretical distribution is chosen, but there is always more or less discrepancy.

The basic idea of the Pearson χ^2 criterion is to discern whether to modify the original predicted value by the degree of difference between the original probability and the statistical distribution actually observed in the experiment, and its basic method is as follows:

The formula for calculating the degree of discrepancy:

$$X_\alpha^2 = \sum_{i=1}^{l} \frac{\left(m_i - np_i\right)^2}{np_i} \tag{13}$$

where m_1 is the cumulative number of the i nd action. n is the cumulative number of individual actions. p_i is then the original predicted probability of each action. For a given significance level α, and the value of X_α^2 at degree of freedom $k = l - r - 1$, such that:

$$P\left(X^2 \geq X_a^2\right) = \alpha \tag{14}$$

where l is the number of divided intervals and r is the number of unknown parameters in the theoretical distribution for which sample observations are to be used.

If the value of X^2 calculated from the latest observations is greater than X_α^2, the original predicted probability should not be accepted at significance level α, otherwise it is accepted. If the original predicted probability is to be rejected, it is to be replaced by a frequency distribution m_i/n, which makes it closer to the actual level.

It should be particularly emphasized that testing hypotheses about prediction rates using the Pearson χ^2 criterion must require that both the number of experiments n and the frequency m_1 of observations within each action interval should be relatively large, usually taking $n \geq 50$ for each $m_i \geq 5$. And in the actual execution, it may be that if an action predicts a probability of 0.0, then it needs to be forced to be set to 0.000001 to prevent Eq. (12) from having an error of dividing by zero.

4. Simulation experiments and analysis of results

4.1. Analysis of multi-intelligence body simulation experiments

In order to verify the effectiveness of the above NashQ algorithm-based multi-intelligence simulation for juvenile delinquency prevention, this paper designs a set of reward and punishment mechanisms to guide the behavior of Agents in the LaserTag experimental environment. For example, if an Agent chooses to comply with legal behaviors, he will get positive rewards, such as social recognition or economic benefits. On the contrary, if an Agent chooses to commit a criminal act, he may be punished, such as legal sanctions or social ostracism. At the beginning of each simulation, two Agents are randomly initialized at the generation point. If one Agent outperforms the other, then it receives an immediate reward value of 1. In total, 1000 rounds of simulation were conducted. Each round contains 1000 iterations and the Nash equilibrium convergence curve of the NashQ algorithm trained 1000 times is shown in Fig. 3.The maximum interval step size of the NashQ algorithm oscillates slightly in the pre-training period but converges to 0 in the later period. This means that the two Agents gradually learn the “skill” of winning rewards, which shows that the convergence of the NashQ algorithm is guaranteed.

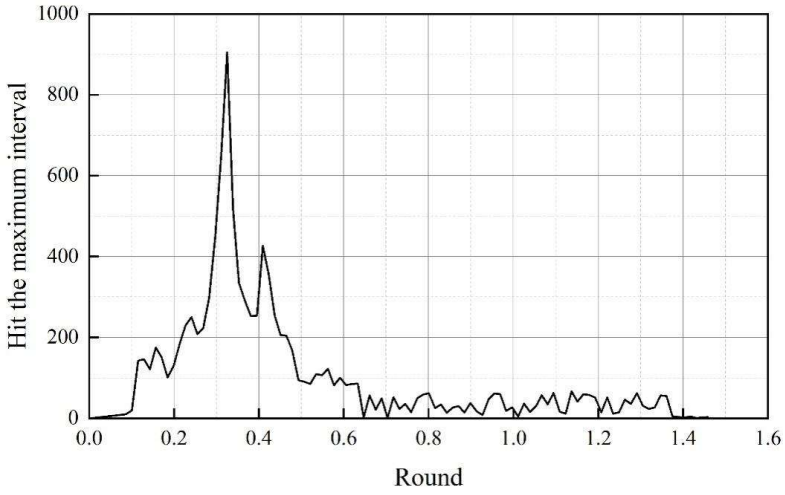


Fig. 3. Nash equilibrium convergence curve

The dual Agent reward curve for the NashQ algorithm is shown in Fig. 4. The round reward obtained by the Agent converges near 300 in the NashQ algorithm. This indicates that in the LaserTag experimental environment, only about 1/3 of the training times of the NashQ algorithm are needed to obtain close to a large boost in the Agent's turn reward. This shows that in multi-agent systems, the NashQ algorithm can boost the total Agent reward value, enhance the Agent's ability to utilize and explore, and dissipate the instability of the environment to some extent.

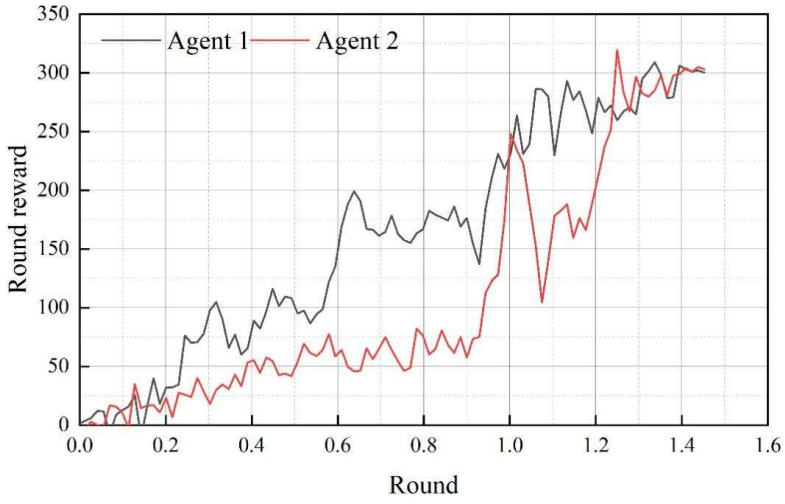


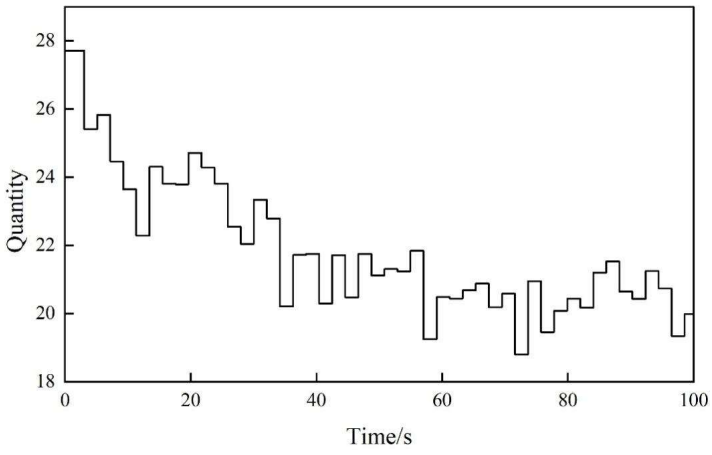
Fig. 4. The NashQ algorithm double Agent reward curve

4.2. Analysis of the effectiveness of prevention under different causes of juvenile delinquency

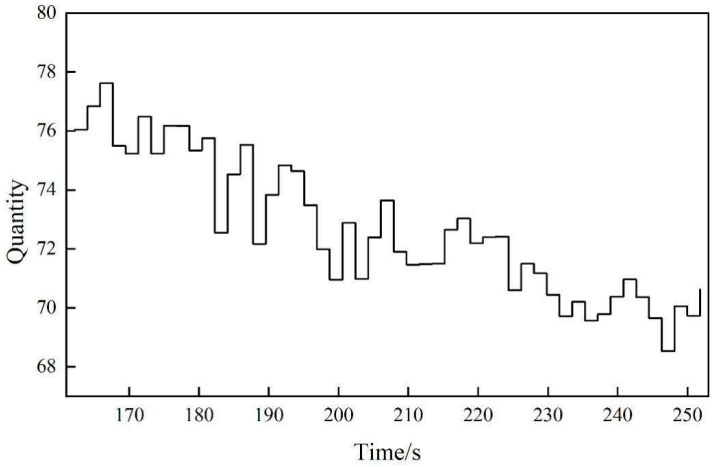
In this paper, we run simulation experiments by varying the initial values of different parameters so as to conduct stability analysis. The size of the inputs of different variables (family, school, and society) is changed sequentially, and the changes in the model outputs are observed to examine the influence of each variable on the stability of the prevention effect.

4.2.1. Validation of the stability of family factors on the effectiveness of prevention. The family environment, including the integrity of the family structure, the degree of harmony in the relationship between family members and the quality of family education, directly affects the behavioral patterns and psychological development of minors. In the simulation process, a stable family environment can provide more positive incentives and support to minors and reduce their tendency to commit crimes,

thus increasing the stability of the multi-intelligence body simulation model. On the contrary, e.g., domestic violence, neglect, or overindulgence may lead to behavioral deviations of minors, increasing the uncertainty and volatility of the model. Therefore, family environment factors are one of the important variables to assess the stability of prevention effects. The prevention effects under different family environment parameters are shown in Figure 5. (a) and (b) represent the prevention effect when the home environment parameter is 50 and 500, respectively. With the increase of home environment parameter, the time for the prevention effect to reach stability is delayed from 22s in Fig. 5(a) to 74s in Fig. 5(b), while the number of Agent individuals also increases. It indicates that the increase in the number of Agent individuals is conducive to improving the overall synergy level of the prevention effect. The reason for this is that as the scale of the dynamic prevention effect increases, the interactions within the prevention effect that make up the underage crime strategy become more complex, which increases the difficulty of coordination between the prevention effects, resulting in a subsequent increase in the cost of coordination of the prevention effect. As a result, it is more difficult to coordinate between nodes within the prevention effect, the stability of the prevention effect as a whole decreases, i.e., the magnitude of change in the prevention effect increases, and the prevention effect requires more time for coordination to reach a stable state. Therefore, a stable family environment is an important strategy for the prevention of juvenile delinquency.



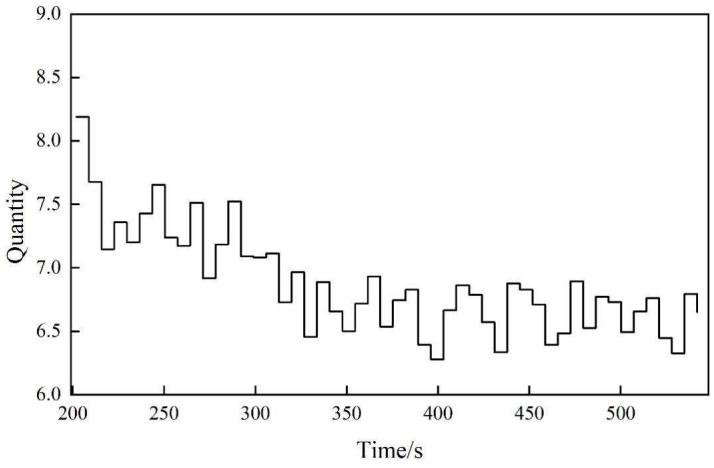
(a) Parameter=50



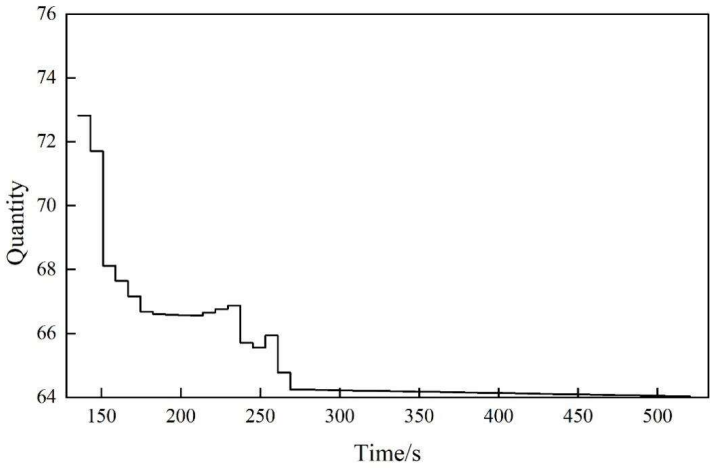
(b) Parameter=500

Fig. 5. The prevention effect of different household environment parameters

4.2.2. Stability validation of prevention effects under school factors. Factors such as the quality of school education, teacher strength, campus culture and peer relationships work together to shape the values and behavioral norms of minors. In the multi-intelligence body simulation model, high-quality educational resources and positive campus culture can promote the development of positive behaviors of minors and reduce the probability of criminal behavior, thus improving the stability of the model. Therefore, the school factor is an important parameter affecting minors' behavioral choices and criminal tendencies in the multi-intelligent body simulation model. The prevention effects under different school environment parameters are shown in Figure 6. (a) and (b) represent the prevention effect when the school environment parameters are 50 and 500, respectively. The time required for the prevention effect to reach equilibrium is shortened from 370s in Fig. 6(a) to 270s in Fig. (b), and the overall stabilization speed of the prevention effect is accelerated. The reason for this is that the prevention effect is related to the quality of education, teacher strength, and campus culture of the school. As the strength of the school increases, the fluctuation amplitude of the prevention effect becomes smaller, and the prevention effect is faster and more stable to achieve equilibrium in the process of evolution.



(a) Parameter=50

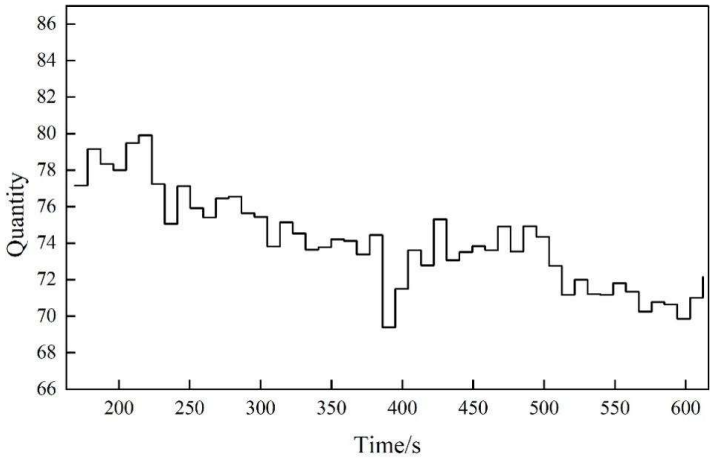


(b) Parameter=500

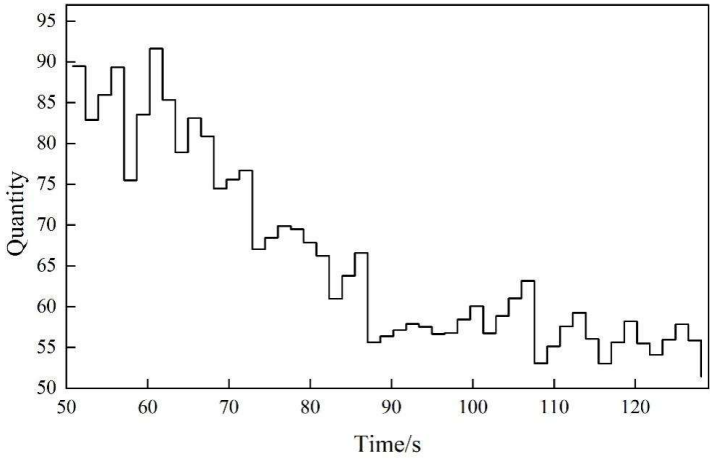
Fig. 6. The prevention effect of different school environment parameters

4.2.3. Stability validation of prevention effects under social factors. The social environment provides minors with a broader context for social interactions and behavioral choices, which has a profound

impact on the stability of the multi-intelligence body simulation model. Factors such as socio-economic status, legal system, employment opportunities and social support system together determine the social pressures and opportunities faced by minors. In the simulation model, a healthy and just social environment can provide minors with more positive incentives and opportunities for success, reduce delinquent behavior, and enhance the stability of the model. In contrast, factors such as social injustice, economic pressure or lack of social support may increase minors' risk of delinquency, leading to fluctuations and uncertainty in the model results. The prevention effects under different social environment parameters are shown in Figure 7. (a) and (b) represent the prevention effect when the social environment parameters are 50 and 500, respectively. The time required for the prevention effect to stabilize is shortened from 520s in Fig. 7(a) to 110s in Fig. 7(b). The improvement of the social environment makes the evolution of the prevention effect stabilize faster.



(a) Parameter=50



(b) Parameter=500

Fig. 7. Prevention effect under different social environment parameters

5. Conclusion

In this paper, a multi-agent reinforcement learning framework is used to construct a multi-intelligent body simulation model for juvenile crime prevention. And the NashQ algorithm is introduced to simulate the strategy choice of minors when facing the temptation of crime. In the simulation environment, the maximum interval step size of the NashQ algorithm is slightly oscillated in the early stage of training, and gradually converges to 0 in the later stage, and only 1/3 of the training times are

needed to achieve the stability of the simulation environment. The NashQ algorithm is used to conduct experimental simulation of crime prevention strategies to explore the stability of prevention effects under different causes of juvenile delinquency. The analysis shows that a good family environment, high quality teaching conditions and a healthy social atmosphere can effectively prevent juvenile delinquency.

Funding

This work was supported by The Humanities and Social Sciences Research Project of the Ministry of Education, "Research on the System of Custody and Reeducation of Minors" (Grant Number: 20YJC820034).

References

[1] Schepers, D. (2017). Causes of the causes of juvenile delinquency: Social disadvantages in the context of Situational Action Theory. European journal of criminology, 14(2), 143-159.

[2] Delcea, C., Fabian, A. M., Radu, C. C., & Dumbravă, D. P. (2019). Juvenile delinquency within the forensic context. Romanian Journal of Legal Medicine, 27(4), 366-372.

[3] Khuda, K. E. (2019). Juvenile delinquency, Its causes and justice system in Bangladesh: A Critical Analysis. Journal of South Asian Studies, 7(3), 111-120.

[4] Rose, S., Ambreen, S., & Fayyaz, W. (2017). Contributing factors of juvenile delinquency among youth of Balochistan. Pakistan Journal of Criminology, 9(3), 165-179.

[5] Trinidad, A., Vozmediano, L., & San-Juan, C. (2020). Environmental factors in juvenile delinquency: A systematic review of the situational perspectives' literature. Reviewing Crime Psychology, 240-266.

[6] Gungea, M., Jaunky, V. C., & Ramesh, V. (2017). Personality traits and juvenile delinquency. International Journal of Conceptions on Management and Social Sciences, 5(1), 42-46.

[7] Zhadan, V. N., Kamalova, G. T., Sadykanova, Z. E., & Karipova, A. Y. (2019). On the problems and directions for the prevention of juvenile delinquency. Journal of Advanced Research in Law and Economics, 10(1 (39)), 401-411.

[8] Young, S., Greer, B., & Church, R. (2017). Juvenile delinquency, welfare, justice and therapeutic interventions: a global perspective. BJPsych bulletin, 41(1), 21-29.

[9] Yan, F. (2023). Research on the Status Quo and Prevention of Juvenile Delinquency from the Perspective of Internet. International Journal of Frontiers in Sociology, 5(9).

[10] Farrington, D. P., Gaffney, H., Lösel, F., & Ttofi, M. M. (2017). Systematic reviews of the effectiveness of developmental prevention programs in reducing delinquency, aggression, and bullying. Aggression and Violent Behavior, 33, 91-106.

[11] van der Laan, A. M., Rokven, J., Weijters, G., & Beerthuizen, M. G. (2021). The drop in juvenile delinquency in the Netherlands: Changes in exposure to risk and protection. Justice quarterly, 38(3), 433-453.

[12] Aw, S., Widiarti, P. W., Setiawan, B., Mustaffa, N., Ali, M. N. S., & Hastasari, C. (2020). Parenting and sharenting communication for preventing juvenile delinquency. Inform, 50(2), 177-186.

[13] Valuiskov, N. V., Bondarenk, L. V., & Arutiunian, A. D. (2017). Juvenile crime: current state and dynamics. J. Pol. & L., 10, 225.

[14] Caskey, T. R., Wasek, J. S., & Franz, A. Y. (2018). Deter and protect: crime modeling with multi-agent learning. Complex & Intelligent Systems, 4(3), 155-169.

[15] Brüngger, R. R., Kadar, C., Cvijikj, I. P., Brüngger, R. R., Kadar, C., & Pletikosa, I. (2016, July). Design of an agent-based model to predict crime (WIP). In SummerSim (p. 55).

[16] Chen, F., & Ren, W. (2019). On the control of multi-agent systems: A survey. Foundations and Trends® in Systems and Control, 6(4), 339-499.

[17] Amirkhani, A., & Barshooi, A. H. (2022). Consensus in multi-agent systems: a review. Artificial Intelligence Review, 55(5), 3897-3935.

[18] Li, Y., & Tan, C. (2019). A survey of the consensus for multi-agent systems. Systems Science & Control Engineering, 7(1), 468-482.

[19] Zheng, Y., Ma, J., & Wang, L. (2017). Consensus of hybrid multi-agent systems. IEEE transactions on neural networks and learning systems, 29(4), 1359-1365.

[20] Sun, Z. (2018). Cooperative coordination and formation control for multi-agent systems. Springer.

[21] Huang Qiqi. (2020). Analysis on the Causes of Juvenile Crimes from the Perspective of Psychology. Academic Journal of Humanities & Social Sciences(10.0).

[22] Masoumeh Rezazadeh Seylab,Mehdi S. Naderi & Gevork B. Gharehpetian. (2024). Dynamic energy management and control of networked microgrids based on load to grid services and incentive-based demand response programs: A multi-agent deep reinforcement learning approach. Sustainable Cities and Society105957-105957.

[23] Chenyang Zhu,Jinyu Zhu,Wen Si,Xueyuan Wang & Fang Wang. (2024). Multi-agent reinforcement learning with synchronized and decomposed reward automaton synthesized from reactive temporal logic. Knowledge-Based Systems112703-112703.