aerospace

Article

# Radar Anti-Jamming Countermeasures Intelligent Decision-Making: A Partially Observable Markov Decision Process Approach

**Huaixi Xing *, Qinghua Xing and Kun Wang ***

Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China
* Correspondence: huaixi_xing@163.com (H.X.); 17765073967@163.com (K.W.)

**Abstract:** Current electronic warfare jammers and radar countermeasures are characterized by dynamism and uncertainty. This paper focuses on a decision-making framework of radar anti-jamming countermeasures. The characteristics and implementation process of radar intelligent anti-jamming systems are analyzed, and a scheduling method for radar anti-jamming action based on the Partially Observable Markov Process (POMDP) is proposed. The sample-based belief distribution is used to reflect the radar's cognition of the environment and describes the uncertainty of the recognition of jamming patterns in the belief state space. The belief state of jamming patterns is updated with Bayesian rules. The reward function is used as the evaluation criterion to select the best anti-jamming strategy, so that the radar is in a low threat state as often as possible. Numerical simulation combines the behavioral prior knowledge base of radars and jammers and obtains the behavioral confrontation benefit matrix from the past experience of experts. The radar controls the output according to the POMDP policy, and dynamically performs the best anti-jamming action according to the change of jamming state. The results show that the POMDP anti-jamming policy is better than the conventional policy. The POMDP approach improves the adaptive anti-jamming capability of the radar and can quickly realize the anti-jamming decision to jammers. This work provides some design ideas for the subsequent development of an intelligent radar.

**Keywords:** anti-jamming countermeasures; POMDP; intelligent radar; electronic warfare

check for updates

**Citation:** Xing, H.; Xing, Q.; Wang, K. Radar Anti-Jamming Countermeasures Intelligent Decision-Making: A Partially Observable Markov Decision Process Approach. *Aerospace* **2023**, *10*, 236. https://doi.org/10.3390/aerospace10030236

Academic Editor: Yan (Rockee) Zhang

Received: 3 January 2023
Revised: 18 February 2023
Accepted: 19 February 2023
Published: 27 February 2023

## 1. Introduction

With the continuous development of electronic warfare, electronic jamming technology to protect the target assault brings severe challenges to the defender, resulting in a serious weakening of radar detection capabilities. Electronic jamming has the characteristics of dynamics, diversity, and complexity, which requires better and better anti-jamming capability of radars. Today, radars have many mature anti-jamming technologies, but commanders are more concerned about how to plan the best anti-jamming measures when they are jammed. The emergence of cognitive radars makes the choice of intelligent decision-making instead of relying on the subjective experience gradually become the mainstream trend [1]. The core of radar intelligent anti-jamming technology is to automatically identify the jamming type and adaptively take effective anti-jamming measures to complete the confrontation [2]. Cognitive radar actively perceives the characteristics of interfering electromagnetic signals in the environment, classifies and identifies the interfering signals, and then selects the corresponding optimal anti-jamming countermeasures. Cognitive radars have the ability to adaptively schedule anti-jamming measures [3].

The electronic countermeasures of jammers and radars are similar to the game process of playing chess [4]. According to the current situation of the chessboard, the two sides choose to place their pieces in ever-changing ways. There are many uncertain factors in the game between jammers and radars, mainly as follows: First, the reliability of radar identification. The jamming types may be incorrectly identified. Second, the mutual restraint relationship between radar anti-jamming measures and jamming types is not one-to-one mapping. Multiple anti-jamming actions can be used to deal with the same jamming pattern, and one anti-jamming measure can also weaken multiple jamming patterns. Therefore, in the complex electromagnetic environment, the choice of anti-jamming action has great uncertainty.

The decision-making principle of anti-jamming action is that the radar takes the best anti-jamming measures to weaken the jamming effect of a certain jamming pattern, so that the detection ability of the radar is improved. Since the jammer signal processor can capture the characteristics of the signal emitted by the radar and adjust the jamming pattern in time according to the radar anti-jamming response, the radar and the jammer are in a dynamic game process [5]. We regard the choice of radar and jammer actions as an uncertain dynamic decision-making problem, and the framework of the decision-making model is stochastic dynamic programming [6]. It has the Markov property, because the conditional probability distribution of its future state depends only on the current state. Markov decision is the optimal decision process of a stochastic dynamic system based on the Markov process theory [7]. Through the study of state space, the set of actions and state transition probability, the future state, and change of the system can be predicted to some extent. In this paper, the goal is to build a sequential decision model in which problems can be modeled as Markov Decision Processes (MDPs) if the state is observable and only its transitions are random. If the state is not fully observable, the problem can be modeled as a partially observable Markov decision process (POMDP) [8]. For our focus, the uncertainty in the choice of radar anti-jamming action can be considered incompletely observable. Therefore, POMDP is a suitable framework for our research content.

The rest of the paper is organised as follows: Related radar anti-jamming decision work is introduced in Section 2. The anti-jamming system design and problem formulation are in Section 3. The proposed POMDP model for intelligent decision-making for radar anti-jamming actions is described in Sections 4 and 5. Simulation results and analysis are given in Section 6. The conclusion is discussed in Section 7.

## 2. Related Work

Jiang et al. [3] considered the radar resource management problem in a cognitive radar for tracking multiple moving targets with jammers. This problem was modeled as a hybrid POMDP-based game model to exploit the statistical characteristics of the moving targets and the competition between the jammers and the radar. They proposed a low complexity gradient optimization algorithm to find the optimal anti-jamming policy of the radar. To separate the hidden real target and establish their accurate trajectories in the presence of deception jammers, Jiang et al. [6] further promoted the existing non-anti-jamming tracking model to the anti-jamming POMDP-based game tracking model. The optimal anti-jamming resource management policy enables the limited resources to be effectively utilized, thus guaranteeing the anti-jamming performance. Similarly, both works also apply the POMDP approach to the radar anti-jamming problem; they studied the optimal policy for radar anti-jamming resource management from the level of signal processing with the goal of improving radar tracking performance in the presence of jammers. However, our work is to design a sequential decision model for radar anti-jamming. During the dynamic confrontation between radar and jammer, the radar is able to select the best anti-jamming measures for different jamming patterns to improve its anti-jamming performance. Joseph Mitola et al. proposed the cognitive radio theory in 1999, which pioneered the research on cognitive radar theory [9]. Chen Wei et al. proposed a combat mode of intelligent networking and coordinated attack, using artificial intelligence methods to enhance the anti-jamming function of radar and seeker and improve the level of intelligence [10]. Liu Meng et al. proposed an intelligent anti-jamming system based on a neural network algorithm, which mainly completed the waveform design from discontinuous orthogonal frequency division multiplexing, a transform domain, and spread spectrum [11]. Wu Xianmeng et al. studied the jamming perception and performance evaluation of low-altitude radars, and gave a technical solution for intelligent anti-jamming [12]. Wang Xin et al. proposed a radar anti-jamming system architecture based on jamming cognition. Through the reconnaissance and identification of jamming, game theory was applied to autonomously select anti-jamming measures [13]. Dr. He Bin [14] modeled and analyzed the game process of cognitive radar anti-jamming. Based on the principle of radar active jamming guidance, he studied the anti-jamming countermeasures of radar radio frequency shielding. Wang Hao [15] and others from Hehai University used Q-Learning and Sarsa methods to enable the radar to update and optimize the anti-jamming policy autonomously. Yuan Quan [16] from Harbin Institute of Technology used the deep Q-network (DQN) algorithm to realize the design of the intelligent radar anti-jamming network. By intelligently identifying the type of jamming, the DQN decision algorithm was applied to intelligently select the best anti-jamming measures. Serkan Ak and Stefan Bruggenwirth investigated the anti-jamming performance of a cognitive radar under a partially observable Markov decision process (POMDP) model. They illuminated the performance metric of probability of being jammed for the radar beyond a conventional signal-to-noise ratio (SNR). This performance metric is analyzed with deep Q-network (DQN) and long short-term memory (LSTM) networks under various uncertainty values [17].

## 3. System Model and Problem Formulation

In this section, we present the system model of radar anti-jamming systems. Then, the sequential decision model of the anti-jamming policy based on POMDP is formulated.

### *3.1. System Structure Design*

We designed a cognitive radar anti-jamming intelligent decision-making system. The POMDP model is used to model the random behavior of adaptively selecting anti-jamming actions when the radar is jammed to achieve intelligent confrontation. For uncertain factors, the POMDP model can be described by probability. The advantage of the model is that it can quantitatively describe the radar's ability to identify the working state of the jammer, and then provide support for the radar's autonomous anti-jamming action decision-making. Figure 1 depicts the structure of the radar intelligent anti-jamming system.

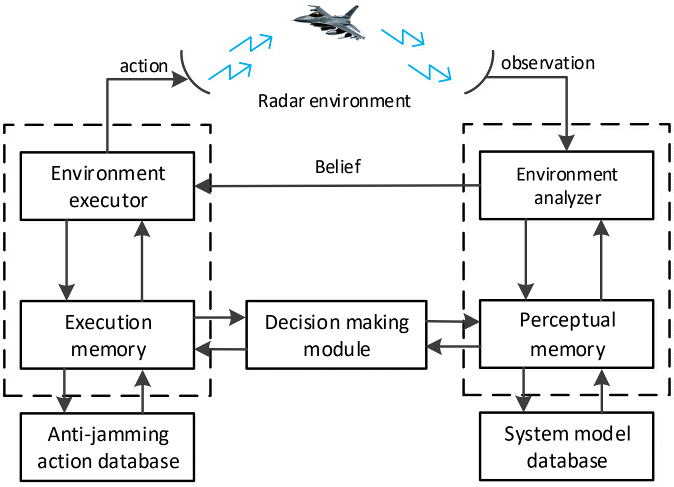


**Figure 1.** The structure of the radar intelligent anti-jamming system.

(1) This system realizes the adaptive closed loop of transmitter→antenna transmission→ space (channel)→antenna reception→receiver→anti-jamming decision→transmitter.

(2) This system has an environmental dynamic database that contains information about the environment and targets of interest, which come from external jamming sources, and the information in the database is constantly updated dynamically.

(3) The adaptive anti-jamming decision-making unit has the function of knowledge-assisted processing.

### 3.2. POMDP Preliminaries

The characteristics of Markov's decision-making process are action space that can be used and the specific actions that will be taken only up to the current state of the system and have nothing to do with past history. The Markov decision process is a special kind of sequential decision problem, which is characterized by the set of actions that can be adopted. The acquired return and transfer probability only depend on the current state and selected actions, and have nothing to do with the past history. This property is called Markov. A POMDP is the generalization of a Markov decision process [18]. It describes a decision maker's interaction with a stochastic system of which the current state is not directly observable. The main part of POMDP is similar to the general sequence decision model; it has a decision cycle, state, action, observation, transition probability, and reward. The model is described by the following elements.

- $S$, the set of system states.
- $A$, the set of actions.
- $O$, the set of observations.
- The observation model denoted by $O(z, s, a)$, i.e., the probability that observation $z$ was made given that the state was $s$ and action $a$ was taken.
- The underlying Markov Chain that models the transitions of the system's state, denoted by $P(s, s', a)$, the probability that the system state transitions from the previous state $s$ to the next state $s'$ after taking action $a$.
- The function $R(s, a)$ defines a real-valued reward when the agent takes action $a$ in state $s$.

As the system's state is not observable exactly, the knowledge about the system is represented by the belief state $b(s)$. This is a probability distribution over the state space based on the internal dynamics of the system. The system starts with an initial belief $b_0$, the actions taken and the observations made. By incorporating information from the action $a$ taken and the observation $z$ received, the updated belief $b_{t+1}(s)$ at time $t+1$ is computed with a Bayesian update [19].

$$b_{t+1}(s) = \frac{O(z, s, a)\sum_{s' \in S} P(s', s, a) b_t(s')}{\sum_{s'' \in S} O(z, s'', a)\sum_{s' \in S} P(s', s'', a) b_t(s')} \tag{1}$$

The combination of an action and an observation induces an immediate reward, depending on the current state, and a future reward, depending on the next state. The value function describes the relation between the immediate reward, future reward, and the belief state. Due to the introduction of the belief state space, the POMDP problem can be transformed into a Markov decision process based on the belief state space. The optimal value function can be transformed into the maximum expected value of the discounted reward at the belief point. Equation (2) is the Bellman optimality equation of POMDP model in the belief state reformulation.

$$V^*(b) = \max_{a \in A}\left[\rho(b, a) + \gamma \sum_{z \in Z} \Pr(b, a, z) V^*(b')\right] \tag{2}$$

Here, $z$ is all possible observations that the agent can receive from the environment, and $b'$ is the belief state of the next time step. $\rho(b, a) = \sum\limits_{s \in S} b(s) R(s, a)$, $\Pr(b, a, z) = \sum\limits_{s' \in S} O(z, s', a) \sum\limits_{s \in S} P(s, s', a) b(s)$. $\Pr(b, a, z)$ indicates the probability that the agent takes an action $a$ and obtains an observation $z$ in the belief state $b$. $\gamma$ is the discount factor. In the decision-making process, the discount factor makes the defined optimization function converge to a finite value. Generally, the smaller the discount factor is, the greater the impact of short-term income is, which makes the agent pay more attention to short-term income when making decisions.

A POMDP policy prescribes the action at a belief. An optimal policy is one that maximizes the value function. This is described by the optimal value function which gives for each belief the maximum value. The optimal policy can be defined by Equation (3).

$$\pi^*(b) = \underset{a \in A}{\operatorname{argmax}}[\rho(b,a) + \gamma \sum_{z \in Z} \Pr(b,a,z) V^*(b')] \tag{3}$$

$b'$ represents the belief state of the next time step. An optimal policy is computationally intractable because it is defined on a continuous belief space. Fortunately, in the finite-horizon case, it is proven that the optimal value function is piece-wise linear and convex (PWLC) [20]. A possible way is to compute an approximation of the optimal value function over a subset of the belief space. The value function $V(b)$ is the combination of piecewise linear value functions in the belief space. These piecewise valued functions represent the segment or hyperplane with the largest value on the corresponding partition of the belief state space. The coefficients of these hyperplanes are called $\alpha$-vectors. Different $\alpha$-vectors correspond to different policies. The value function under any belief state can be obtained by substituting the belief state into the hyperplane equation. The optimal value function can be written as Equation (4).

$$V^*(b) = \max_k \left\{ \sum_S b(s) \alpha^k(s) \right\} \tag{4}$$

It is noteworthy that Equation (4) is not valid generally for infinite-horizon POMDP, as the finiteness of the set of $\alpha$-vectors does not generally extend from the finite-horizon case to the infinite-horizon case [21]. In the infinite-horizon case, the approximate value of the optimal value function can be obtained using a finite set of vectors. For a certain finite set $\left\{ \alpha^k \right\}$ of so-termed $\alpha$-vectors, each $\alpha$-vector has a corresponding action. The best action to be taken under belief state can be determined by an inner product operation. These POMDP can be solved in an approximation algorithm. An excellent example of the approximation algorithm is the Successive Approximations of the Reachable Space under Optimal Policies (SARSOP).

### *3.3. Point-Based Value Iterative Approximation on SARSOP Algorithm*

Due to the limitation of computational complexity, solving the POMDP problem accurately has always been a difficult problem restricting the development and application of POMDP theory. Many scholars have successively developed many efficient POMDP approximate solving algorithms. H. Kurniawati et al. [22] proposed the Successive Approximations of the Reachable Space under Optimal Policies (SARSOP) algorithm on the basis of previous research. The SARSOP algorithm utilizes the concept of optimally reachable belief space $\Re^*(b_0)$ to improve the computational efficiency of POMDP planning. $\Re^*(b_0)$ is defined as the set of belief states that can be reached from the initial belief $b_0$ by taking an optimal policy. The belief space can be formed as a belief state tree $T_R$ with the initial belief state $b_0$ as the root node. Each node represents a reachable belief state, and all nodes in $T_R$ form the complete belief space. The core of the SARSOP algorithm is the idea of limiting the boundary. Heuristic information and online learning techniques make as many sampled belief states as possible belong to the belief space $\Re^*(b_0)$. The algorithm mainly includes three steps: Sampling, Backup, and Pruning. In order to reduce the amount of computation, the SARSOP algorithm only updates the value function on some belief states, so we need to first sample the belief states according to the belief tree $T_R$ and the initial $\alpha$-vector set. The purpose of performing the backup operation on the belief state $b$ selected by $T_R$ is to sort out the information of sub-nodes and feed it back to $b$. The backup operation is to input a specific belief point and the value function of step $t-1$, and output the optimal vector of point $b$ in the value function of step $t$. For all $a$ and $z$, the next belief state $b'$ can be obtained according to Equation (1), and its corresponding optimal $\alpha$-vector can be obtained from Equation (5).

$$\alpha^{a,z} = \underset{\alpha \in \Gamma}{\operatorname{argmax}} \alpha \cdot b' \tag{5}$$

The value function of all $a$ and $s$ can be calculated by Equation (6).

$$\alpha^{a}(s) = R(s,a) + \gamma \sum_{z \in Z, s' \in S} O(z,s',a) P(s,s',a) \alpha^{a,z}(s') \tag{6}$$

The action $a'$ corresponding to the optimal vector $\alpha^{a'}$ is determined by Equation (7).

$$a' = \underset{a \in A}{\operatorname{argmax}} \alpha^{a} \cdot b \tag{7}$$

Add the optimal vector $\alpha^{a'}$ into the vector set $\Gamma$. The SARSOP algorithm adopts the $\delta$–neighborhood method [20] to cut the branches and vectors corresponding to the suboptimal actions. Furthermore, the algorithm can further reduce the number of vectors in $\Gamma$ by concentrating the sampled belief states nearby.

The introduction of the gap termination condition in SARSOP means that the expected gap $\varepsilon$ is set at the upper limit $\overline{V}$ and lower limit $\underline{V}$ of the value function at the root of $T_R$. By selecting the action and observation of the maximum upper limit on each node, a path that minimizes the root gap is obtained. When the root gap is less than $\varepsilon$ or the time limit is reached, the sampling path is terminated. The pseudo code of SARSOP algorithm is as follows (Algorithm 1).

**Algorithm 1** SARSOP algorithm

```
Initialize $\alpha$ vector set $\Gamma$, initialize the lower bound $\underline{V}$ and upper bound $\bar{V}$ of the value function.
Add initial belief point $b_0$ as the root node of the tree $T_R$
Repeat
    SAMPLE($T_R$,$\Gamma$).
    Select child nodes from $T_R$, for each selected node $b$, BACKUP($T_R$,$\Gamma$,$b$).
    PRUNE($T_R$,$\Gamma$).
Until the root gap is less than $\varepsilon$ or the time limit is reached.
Return $\Gamma$.
```

## 4. Radar Anti-Jamming Action Decision Model

### *4.1. Principles of Radar–Environment Interaction*

Anti-jamming action scheduling is based on identifying, classifying, and building a database, adaptively selecting anti-jamming actions according to the type of jamming, and optimizing anti-jamming actions according to the anti-jamming performance. The same jamming type will have multiple anti-jamming means, and the same jamming countermeasure can also fight against multiple types of jamming. In different jamming environments, different anti-jamming means will have different effects. In addition, the actual battlefield jamming to radar presents a complex and changeable situation, which makes the selection of jamming action extremely complicated.

The radar system contains a dynamic knowledge base for storing, scheduling, and updating various types of prior information. Through the continuous optimization of the dynamic knowledge base, we can find anti-jamming action sets that can deal with a variety of complex scenarios. The cognitive radar anti-jamming system mainly selects anti-jamming actions based on the idea of "game theory". The radar system has complete judgment criteria as to whether the anti-jamming action that can be provided when the jamming is encountered meets the requirements. It is necessary to evaluate the anti-jamming effect, adopt actions with better effects, and the evaluation results provide support for the selection of anti-jamming actions, forming closed-loop feedback.

Knowledge bases and rules need to be preset to describe process tasks and task requirements for a certain period of time. The system provides a set of optimal policies calculated based on the action set and the state set. The radar chooses the best action based on the observations. The POMDP model of radar anti-jamming must ensure that the action of the radar at the next moment is only related to the current jamming signal type. Figure 2 shows the model structure and flow of POMDP for radar anti-jamming.

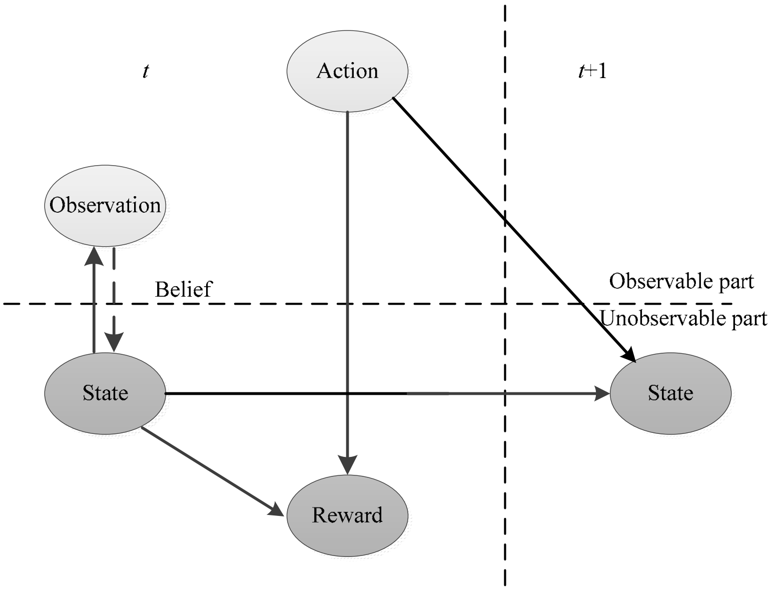


**Figure 2.** Implementation of the POMDP model.

### *4.2. Jamming Types and Belief States*

Many scholars have summarized the benefits of different jamming policies to deal with radar anti-jamming through exploration and research. Reference [23] lists 10 kinds of jamming measures and 9 kinds of radar anti-jamming measures according to the application of electronic countermeasures as in Table 1. It is assumed that 10 jamming measures of the jammer constitute a state set, denoted as $S = \{s_1, \ldots, s_{10}\}$, and 9 anti-jamming measures constitute an action set, denoted as $A = \{a_1, \ldots, a_9\}$. The adversarial jamming effect are shown in Table 2. The smaller the value, the better the anti-jamming effect, otherwise, the worse the anti-jamming effect.

**Table 1.** A library of jamming and anti-jamming policies.

| **Jamming Pattern** | | **Anti-Jamming Action** | |
|---|---|---|---|
| $s_1$ | broadband-noise | $a_1$ | frequency agility |
| $s_2$ | sweep-spot jamming | $a_2$ | pulse stagger |
| $s_3$ | wide pulse | $a_3$ | phase encoding |
| $s_4$ | narrow pulse | $a_4$ | MTD |
| $s_5$ | aiming jamming | $a_5$ | doppler processing |
| $s_6$ | sidelobe jamming | $a_6$ | PRF jitter |
| $s_7$ | main lobe jamming | $a_7$ | single pulse |
| $s_8$ | range gate pull off | $a_8$ | fast AGC |
| $s_9$ | velocity gate pull off | $a_9$ | constant false alarm |
| $s_{10}$ | angle deception jamming | | |

**Table 2.** The benefit matrix of jamming countermeasure radar anti-jamming means.

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ | $s_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | –3 | –10 | –10 | –5 | –10 | –5 | –5 | 3 | 3 | 3 |
| $a_2$ | 3 | –2 | –5 | –5 | –2 | 3 | 3 | –3 | 3 | 3 |
| $a_3$ | –3 | –3 | –5 | –5 | –5 | –5 | –5 | –5 | 3 | 3 |
| $a_4$ | 4 | –2 | –2 | –2 | –10 | –5 | –5 | 4 | 4 | –4 |
| $a_5$ | 5 | –2 | –2 | –2 | –10 | 5 | 5 | –5 | –5 | –5 |
| $a_6$ | 3 | 3 | –2 | –2 | 3 | 3 | 3 | –3 | –3 | 3 |
| $a_7$ | 4 | 4 | 4 | 4 | 4 | 4 | –5 | 4 | 4 | –10 |
| $a_8$ | 2 | 2 | –4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| $a_9$ | 3 | 3 | –2 | 3 | -3 | 2 | 2 | 2 | 2 | 2 |

### *4.3. Radar Anti-Jamming Action and Jamming State Transfer*

Based on the radar countermeasure benefit matrix, the basic principle of state transition is as follows: under the premise of the current anti-jamming policy adopted by the radar, the score of the transferred jamming measure is not less than the score of the current jamming measure. According to the principle of maximum entropy, it is assumed that the state transition probability obeys a uniform distribution in the transferable direction. For example, after executing an action $a$, state $s_1$ can transition to state $s_2$ with a certain probability. $s_1 \rightarrow s_2$ is a transferable direction under action $a$. For the action $s_i$, state $s_i$ can only transition to state $s'$ if the jamming effect of the next state s is better than that of the current state s. Otherwise, the transition probability is 0, as shown in Equation (8). $v(s_i, a_j)$ is the score of current state $s_i$ against anti-jamming action $a_j$. $v(s', a_j)$ is the score of next state $s'$ against anti-jamming action $s_i$.

$$\begin{cases} P(s_i, s', a_j) > 0, & v(s', a_j) \geq v(s_i, a_j) \\ P(s_i, s', a_j) = 0, & v(s', a_j) < v(s_i, a_j) \end{cases} \tag{8}$$

### *4.4. Observations and Observed Probabilities*

Jamming identification plays a very important role in the process of intelligent radar countermeasure. The POMDP approach has the ability of synchronous learning and confrontation by quickly identifying interference features and reasoning interference types. Suppose there are nine groups of time-frequency domain characteristic parameters $\{z_1 \sim z_9\}$ describing the interference type. We also assume that the observation probabilities are already well learned. According to the characteristic samples collected from 10 interference states, the observation probability database of characteristic parameters is constructed, as shown in Table 3.

**Table 3.** Jamming pattern characteristic parameter observation value and observation probability.

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ | $s_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $z_1$ | 0 | 0.11 | 0.11 | 0 | 0.34 | 0 | 0 | 0 | 0 | 0.11 |
| $z_2$ | 0 | 0 | 0.22 | 0.34 | 0.11 | 0.34 | 0.45 | 0.22 | 0.11 | 0.11 |
| $z_3$ | 0 | 0 | 0.11 | 0 | 0 | 0 | 0 | 0 | 0 | 0.11 |
| $z_4$ | 0.22 | 0.11 | 0 | 0 | 0.11 | 0 | 0 | 0.22 | 0.11 | 0 |
| $z_5$ | 0 | 0.34 | 0.45 | 0.34 | 0.11 | 0 | 0 | 0 | 0 | 0 |
| $z_6$ | 0.11 | 0.11 | 0 | 0.11 | 0.11 | 0.22 | 0.22 | 0.22 | 0.22 | 0.22 |
| $z_7$ | 0.34 | 0.22 | 0 | 0.11 | 0.11 | 0.22 | 0.22 | 0.11 | 0.34 | 0.45 |
| $z_8$ | 0.22 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0 | 0.22 | 0.22 | 0 |
| $z_9$ | 0.11 | 0 | 0 | 0 | 0 | 0.11 | 0.11 | 0 | 0 | 0 |

### 4.5. Reward Feedback

The purpose of radar anti-jamming is to weaken the jamming effect or force the jamming state to a less threatening state. To determine the reward function, we define the jamming threat level $\omega(s)$. The efficiency of a radar anti-jamming system is evaluated by comparing the threat level before and after a state transition. In general, if the threat level $\omega(s)$ rises after a state transition, the reward increases. We adopt a subtle way of describing the effectiveness of any one anti-jamming measure in terms of a symbolic function $sign(\bullet)$. If $sign(v(s,a)) = 1$, the qualitative evaluation of the effect of action $a$ is negative. Conversely, if $sign(v(s,a)) = -1$, the qualitative evaluation of the action $a$ is positive. The threat level $\omega(s_i)$ and $\omega(s_j)$ can be compared by Equation (9). Here, If $\sum\limits_{a\in A} sign(v(s_i,a)) = \sum\limits_{a\in A} sign\big(v\big(s_j,a\big)\big)$, the threat level ranking needs to be further determined by calculating the sum of the scores of state $s_i$ and $s_j$ against all actions $a$, as in Equation (10).

$$\begin{cases} \omega(s_i) > \omega\big(s_j\big), \ \sum\limits_{a\in A} sign(v(s_i,a)) > \sum\limits_{a\in A} sign\big(v\big(s_j,a\big)\big) \text{ ①} \\ \omega(s_i) < \omega\big(s_j\big), \ \sum\limits_{a\in A} sign(v(s_i,a)) < \sum\limits_{a\in A} sign\big(v\big(s_j,a\big)\big) \text{ ②} \end{cases} \tag{9}$$

$$\begin{cases} \omega(s_i) > \omega\big(s_j\big), \ \sum\limits_{a\in A} v(s_i,a) > \sum\limits_{a\in A} v\big(s_j,a\big) \text{ ③} \\ \omega(s_i) < \omega\big(s_j\big), \ \sum\limits_{a\in A} v(s_i,a) < \sum\limits_{a\in A} v\big(s_j,a\big) \text{ ④} \\ \omega(s_i) = \omega\big(s_j\big), \ \sum\limits_{a\in A} v(s_i,a) = \sum\limits_{a\in A} v\big(s_j,a\big) \text{ ⑤} \end{cases} \tag{10}$$

According to Equations (9) and (10), it can be obtained that the state set $S$ is arranged into $\omega(s_3) < \omega(s_4) < \omega(s_5) < \omega(s_2) < \omega(s_6) = \omega(s_7) < \omega(s_8) < \omega(s_{10}) < \omega(s_9) < \omega(s_1)$ in the ascending order of the degree of jamming threat.

The reward function obtained by the radar anti-jamming policy is defined as Equation (11).

$$R\big(s',s,a\big) = \begin{cases} 1, \ \omega(s') < \omega(s) \\ 0, \ \omega(s') = \omega(s) \\ -1, \omega(s') > \omega(s) \end{cases} \tag{11}$$

Equation (11) shows that if the threat level of state $s'$ is lower than the previous state $s$, the feedback reward is +1, otherwise, the feedback reward is $-1$, and if the threat level of the two states is the same, the feedback reward is 0.

## 5. Radar Intelligent Anti-Jamming Planning Decision Making Process Based on POMDP

Considering that the game theory-based decision-making of jammers and radars has high real-time requirements, this paper adopts the POMDP offline planning method to select the optimal policy. First, a large amount of preprocessing time is invested in the search phase to generate the radar anti-jamming policy set on the entire belief state space. Then, the optimal policy is selected according to the anti-jamming gain in the policy execution stage. The algorithm flow and the pseudo code are described as follows (Algorithm 2).

Step 1: Input the parameters of the radar anti-jamming policy scheduling model into the SARSOP algorithm, set the expected gap $\varepsilon$ or time limit as the algorithm termination condition, and obtain the $\alpha$-vector set.

Step 2: Initialize the jamming state $s_0$ and the initial belief state $b_0$.

Step 3: Infer the jamming type according to the observed value, and update the belief state $b$ according to Equation (1).

Step 4: Perform an inner product operation on each vector in $\Gamma$ and $b$, and select the optimal anti-jamming action from the policy set according to the $\alpha$-vector corresponding to the maximum value.

Step 5: Affected by the anti-jamming action, the jamming pattern is transferred according to the state transition probability, and new observations are generated.

Step 6: Determine whether the maximum time step $t_{\max}$ is reached; otherwise, return to Step 3.

**Algorithm 2** Offline planning of anti-jamming action based on POMDP

```
Input: A POMDP <$S$,$A$,$O$,$P$,$R$>.
Output: A near-optimal policy $\pi$
Let $V(s) \leftarrow 0$ for all $s \in S$.
repeat
    for each $s \in S$. do
        for each $a \in A$. do
        $V^*(b) = \max_{a \in A}\left[\rho(b,a) + \gamma \sum_{z \in Z} \Pr(b,a,z)V^*(b')\right]$
         end
        $\pi^*(b) = \operatorname{argmax}_{a \in A}[\rho(b,a) + \gamma \sum_{z \in Z} \Pr(b,a,z)V^*(b')]$
        $V^*(b) = \max_k\left\{\sum_S b(s)\alpha^k(s)\right\}$
        update $b(s)$
      end
until $t = t_{\max}$
return $\pi*$
```

## 6. Simulations and Discussions

In this section, the performance of the radar anti-jamming policy based on POMDP is verified by simulation in MATLAB.

Parameter settings are as follows: discount factor $\gamma = 0.8$, the initial belief state obeys a uniform distribution, and the SARSOP algorithm stops when the root gap reaches the expected gap $\varepsilon$; the maximum time step $t_{\max} = 1000$.

The jamming state is dynamically shifted by the utility of radar anti-jamming actions, and observations are generated. The radar updates the belief state and infers the jamming type based on the observations and adaptively selects the best anti-jamming action. The jammer initialization state is broadband noise jamming pattern $s_1$. In the simulation, three anti-jamming action policies are compared with the action policies solved by the POMDP model, including greedy policy, random policy, and fixed policy.

The greedy policy, also labeled as "myopic" policy, determines a state-specific action to the state that has the maximum likelihood of transmitting the current observation. The random policy maps the belief state to a random action, and each action has the same opportunity to be chosen. A fixed policy means that the radar always takes an identical anti-jamming action regardless of how the jamming state changes. The constant false alarm processing is used as a fixed anti-jamming measure for the radar. Figures 3 and 4 shows the update of jamming states and observations under the four policies, and Figure 5 shows an anti-jamming action taken by the radar at each time step. The horizontal axis is the iteration time step from 0 to 50 times and the vertical axis is the sequence number of the state, observations, and action, respectively. The results intuitively reflect the dynamic game confrontation process between the jammer and radar. In order to compare the pros and cons of the four policies, the average cumulative reward value under the four policies is recorded in Figure 6, which means the cumulative reward before the current time step. The value of the time step was varied from 0 to 1000. It can be seen that the reward of the action policy planned by POMDP is significantly better than the greedy policy, random policy, and fixed policy, indicating that the POMDP policy can take more effective anti-jamming measures against the enemy's jamming pattern.

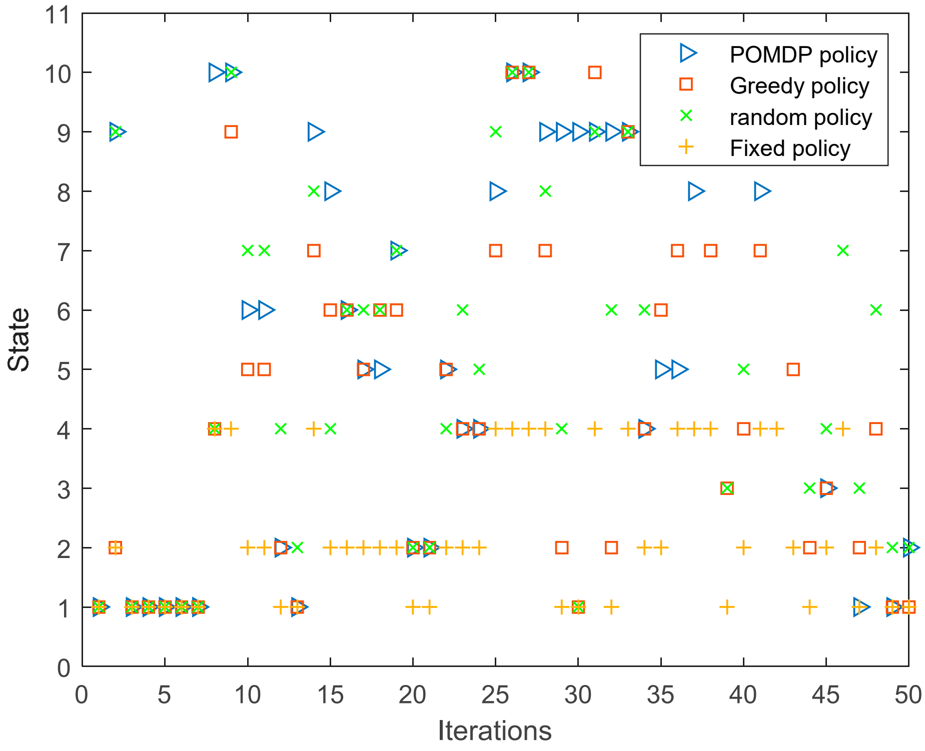


**Figure 3.** The jamming states at each step.

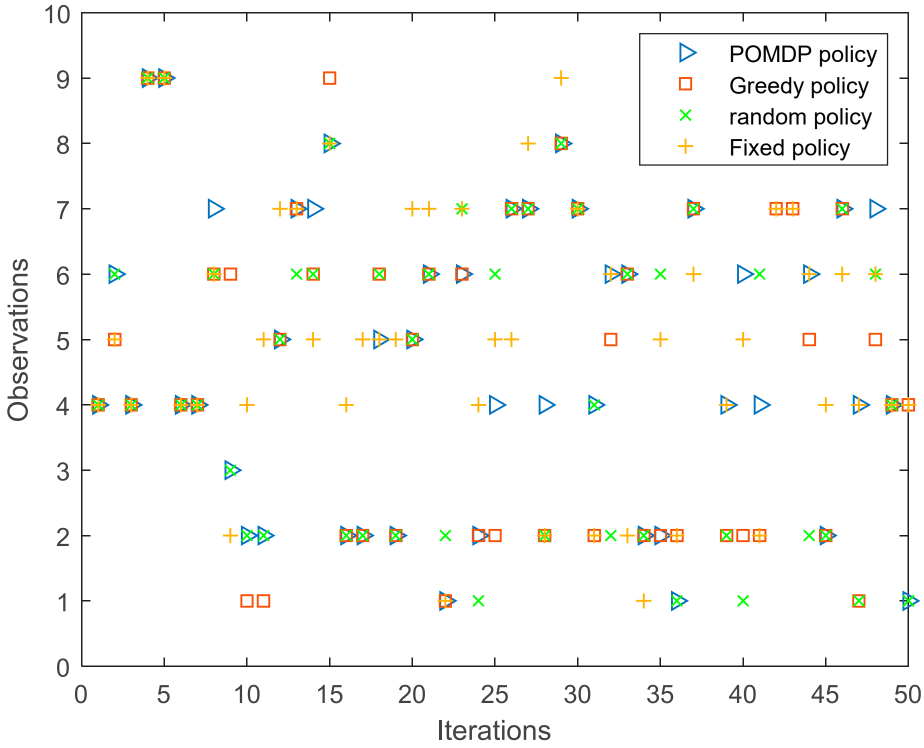


**Figure 4.** The observations at each step.

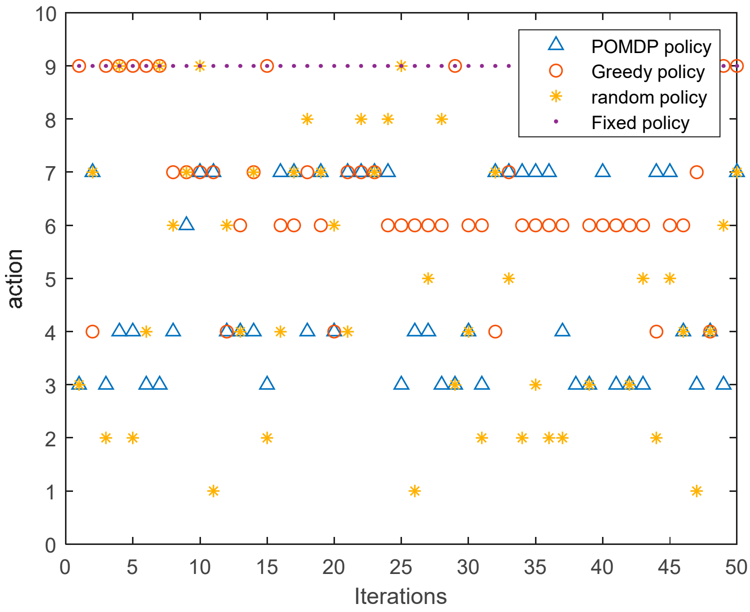


**Figure 5.** The anti-jamming actions at each step.

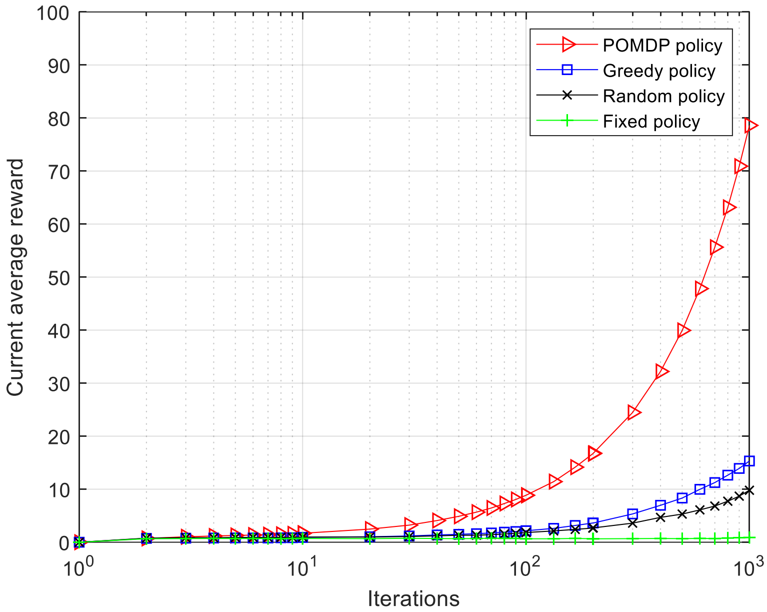
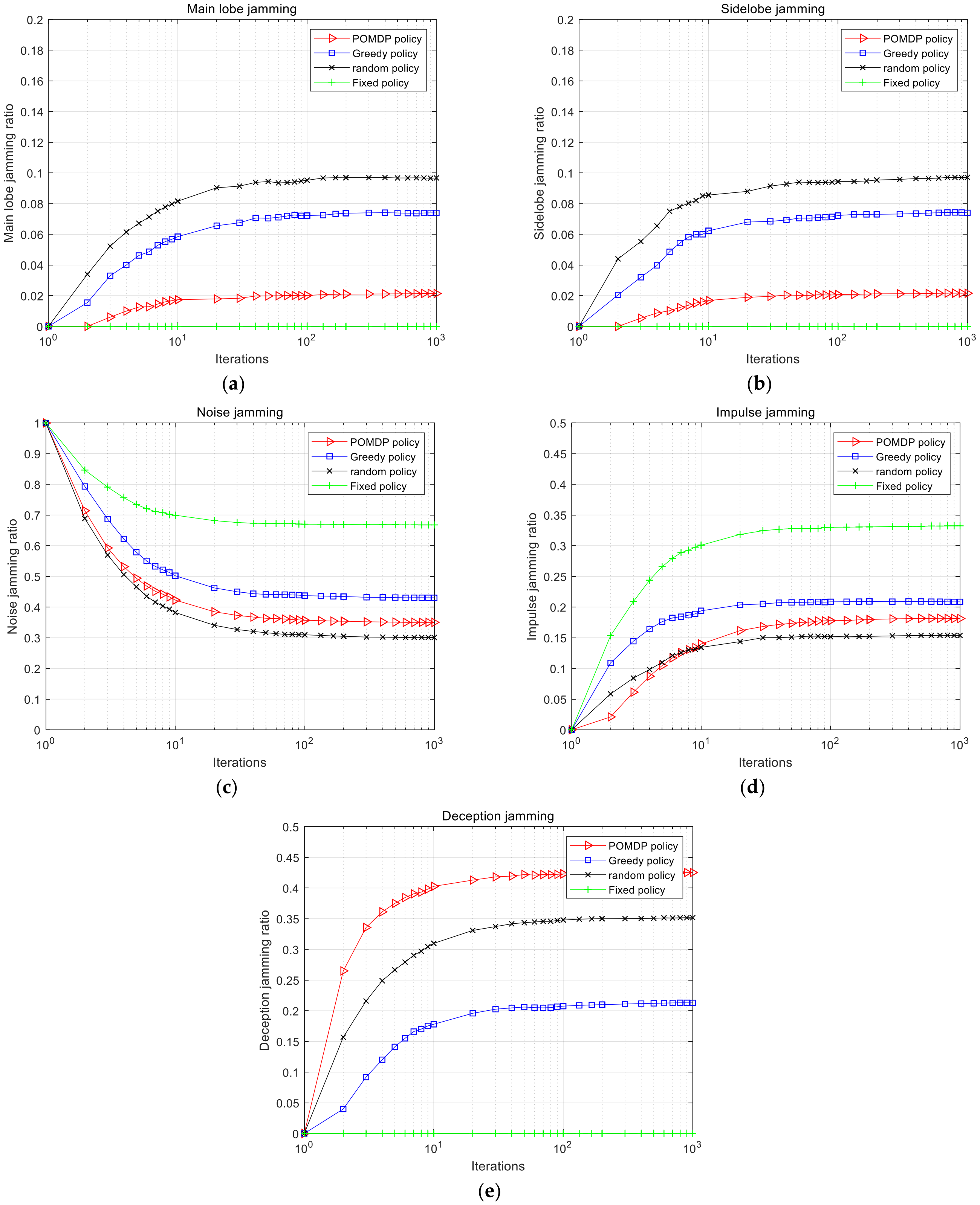


**Figure 6.** The average reward accumulation per time step.

The jamming patterns are divided into five types according to their attributes, namely main lobe jamming ($s_7$), side lobe jamming ($s_6$), noise jamming ($s_1$, $s_2$, $s_5$), impulse jamming ($s_3$, $s_4$), and deception jamming ($s_8$, $s_9$, $s_{10}$). We defined the jamming state occupancy and performed 1000 Monte Carlo experiments to count the average occupancy of different jamming types, as shown in Figure 7. The vertical axis indicates the average occupancy of each jamming type. The average occupancy is defined as follows: the occurrences of a jamming type in all Monte Carlo trails.



(**a**) (**b**) (**c**) (**d**) (**e**)

**Figure 7.** The average occupancy ratio of different jamming types under four policies.

It can be seen that with the transition of states, the average occupancy of the jamming states of the four policies tends to be stable. Except for the fixed policy, the anti-jamming action set generated by the POMDP policy has the highest average occupancy rate for deception jamming, and the lowest average occupancy rate for main lobe jamming and sidelobe jamming. The random policy has the highest average occupancy rate for main lobe and sidelobe jamming, and the lowest average occupancy rate for noise jamming and impulse jamming. Only noise jamming and impulse jamming will occur with a fixed policy.

The jamming states, observations and radar anti-jamming action scheduling are shown in Figures 8–10. We recorded the accumulation of rewards when executing the optimal policy, as shown in Figure 11.

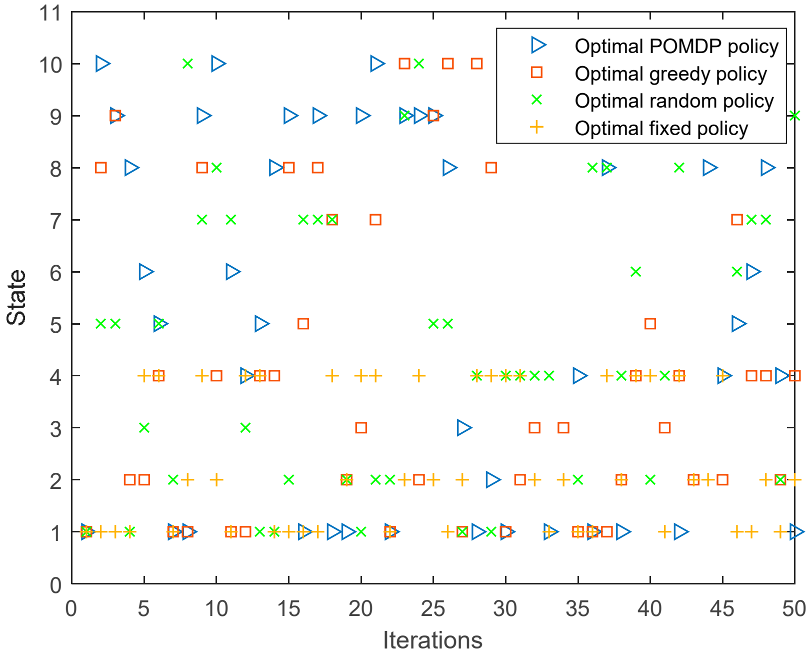


**Figure 8.** The jamming states at each step (optimal policy).

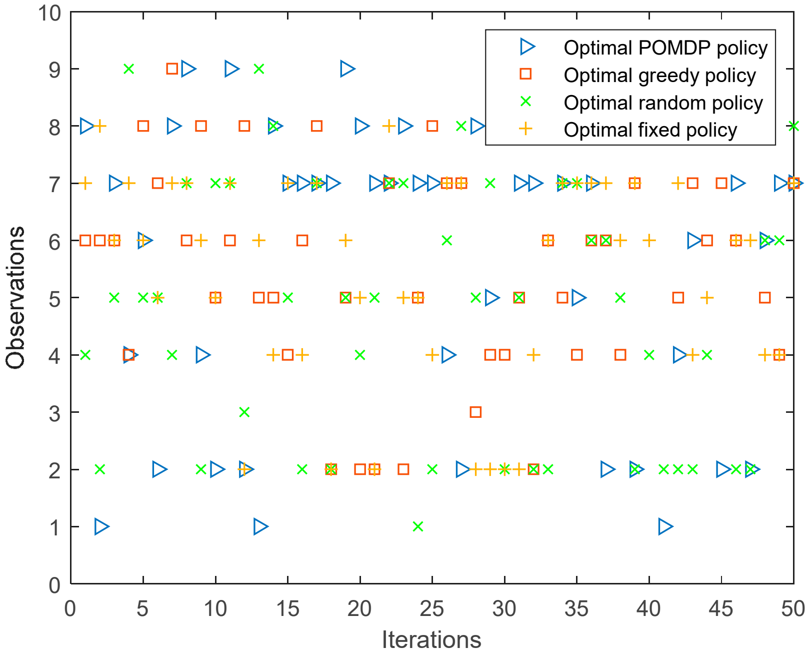


**Figure 9.** The observations at each step (optimal policy).

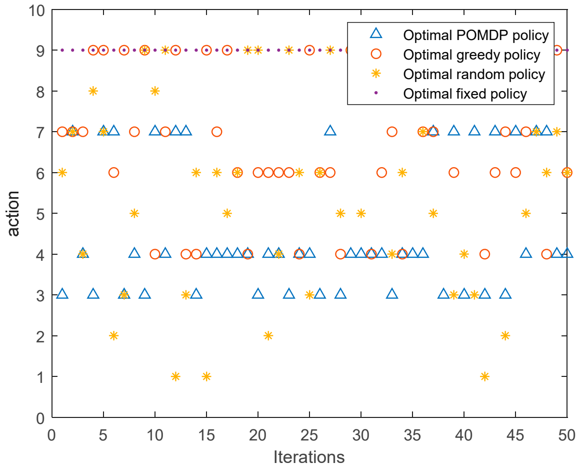


**Figure 10.** The anti-jamming actions at each step (optimal policy).

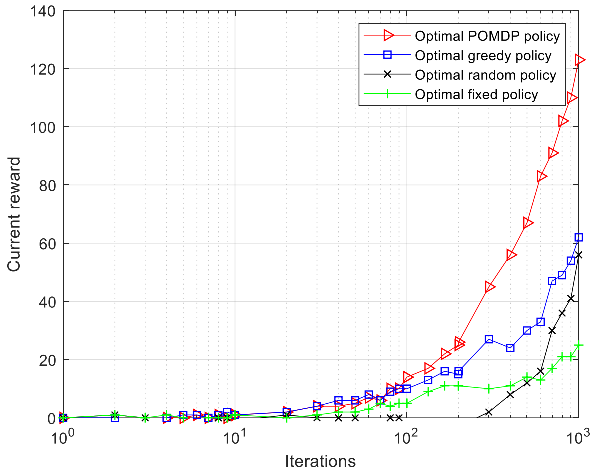
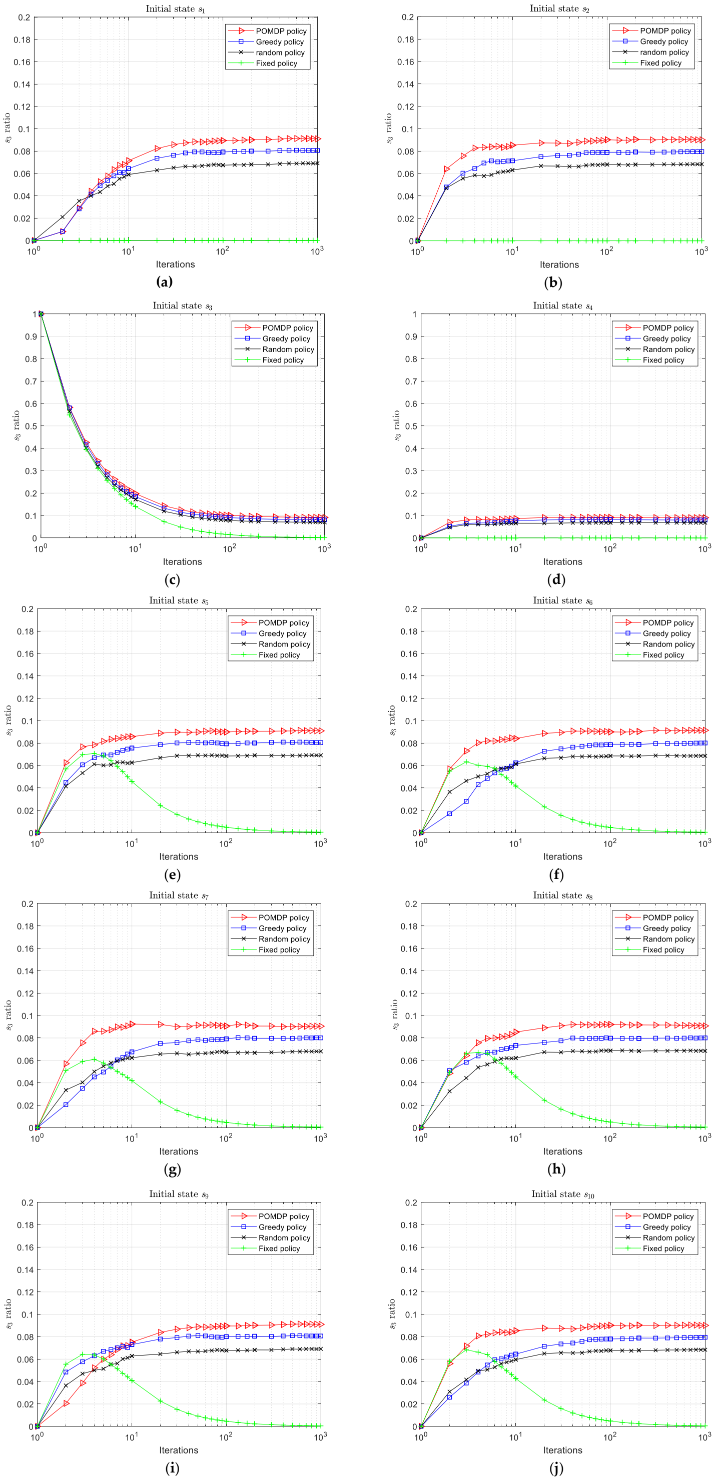


**Figure 11.** The average reward accumulation per time step (optimal policy).

According to the principle of state transition, for the same anti-jamming action, the jamming score of the next state is not inferior to the current state. Anti-jamming effect after state transition will not be improved. After a multi-step transition of state, even the anti-jamming effect is negative. Thus, the long-term reward accumulation of fixed policy is the lowest. The reward of the random policy has strong randomness—each action has the same probability of being selected. The random policy has the lowest short-term reward, but the reward accumulation is higher than that of the fixed policy when the time step reaches about 550. The early-stage greedy policy rewards higher than the POMDP policy. The greedy policy only pays attention to the reward of the current time step, while the POMDP policy pays more attention to the long-term reward, so the former has a higher short-term reward, but at the end of the confrontation, the latter has the highest reward, as shown in Figure 11. In general, the action plan of the POMDP policy has the best anti-jamming effect.

The main goal of the radar's anti-jamming action is to force the jammer to transfer from a high-threat state to a low-threat one, so that the radar can more fully exert its combat power in a safe environment. In order to verify the efficiency of the cognitive radar anti-jamming model based on POMDP in this paper, we changed the initial jamming state settings and counted the average occupancy rate of the jamming state $s_3$ when four policies were adopted, as shown in Figure 12. The vertical axis represents the average occupancy of the lowest threat state $s_3$. The POMDP policy is most effective for forcing the jammer to transfer from a high-threat state to a low-threat one. Meanwhile, the proposed method is insensitive to the initial state. Changing the initial state hardly affects the reward outcome and the average occupancy of the jamming state $s_3$. Table 4 counts the maximum reward and average reward after 1000 steps of anti-jamming action with four policies. Obviously, the reward of the POMDP policy is much higher than the other three policies, while the fixed policy returns the lowest reward and has the worst anti-jamming effect. The average occupancy of jamming state $s_3$ with the lowest threat level is approximately 0.09, 0.08, 0.67, and 0 under the four policies, respectively.



**(a)** **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)** **(i)** **(j)**

**Figure 12.** The average reward accumulation per time step for different initial states.

**Table 4.** Rewards for anti-jamming actions with different initial states.

| Maximum Total Reward | | | | Average Total Reward | | | |
|---|---|---|---|---|---|---|---|
| **POMDP** | **Greedy** | **Random** | **Fixed** | **POMDP** | **Greedy** | **Random** | **Fixed** |
| 122 | 57 | 57 | 28 | 78.611 | 15.319 | 9.801 | 0.894 |
| 125 | 60 | 56 | 33 | 76.081 | 13.477 | 8.727 | −0.451 |
| 120 | 64 | 52 | 23 | 76.005 | 13.806 | 7.348 | −1.402 |
| 122 | 53 | 51 | 27 | 76.305 | 13.888 | 9.124 | −0.852 |
| 122 | 65 | 53 | 10 | 76.597 | 13.911 | 8.116 | −0.471 |
| 121 | 57 | 66 | 26 | 76.257 | 14.256 | 8.604 | −0.057 |
| 121 | 62 | 55 | 28 | 76.919 | 14.673 | 9.015 | −0.518 |
| 121 | 62 | 55 | 28 | 77.141 | 14.882 | 9.274 | −0.101 |
| 122 | 57 | 56 | 27 | 78.164 | 15.227 | 9.527 | 0.71 |
| 125 | 62 | 57 | 32 | 76.655 | 14.179 | 9.371 | −0.07 |

## 7. Conclusions

In order to improve the adaptive anti-jamming capability of radars, a cognitive radar intelligent anti-jamming decision-making system model is designed. The decision-making of the jammer–radar confrontation is realized by applying the POMDP theory, which is different from greedy methods based on expert systems, artificial blind, or fixed selection methods. The system model simulates the Observation, Orientation, Decision, and Action (OODA) closed-loop loop of the anti-jamming system and obtains feedback through the reward function. Considering the uncertain factors of radar identification of jamming types, the jamming types are inferred according to observation so as to select the best action measures, which greatly improves the intelligence of the radar system's anti-jamming decision-making. The average occupancy of the state of the lowest threat level for POMDP policy is the highest in all policies. Compared to the greedy policy and random policy, the occupancy of the lowest threat level state under the POMDP policy increased by approximately 12.5% and 34.3%, respectively. The reward of the POMDP policy is much higher than the other three policies. Therefore, the anti-jamming intelligent decision-making radar system can generate countermeasures adaptively based on the POMDP model, thereby significantly reducing jamming threat to radars and improving the survivability and combat capability of radars.

**Author Contributions:** Conceptualization, H.X.; methodology, H.X.; software, H.X.; validation, H.X. and Q.X.; formal analysis, H.X. and K.W.; investigation, K.W.; resources, H.X.; writing—original draft, H.X. and K.W.; writing—review and editing, Q.X.; supervision, Q.X.; funding acquisition, Q.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by [The National Natural Science Foundation of China] grant number [71771216].

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors are very grateful of the referees for their valuable remarks, which improves the presentation of the paper. This work was supported by the National Natural Science Foundation of China (71771216).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gao, N.; Qin, Z.; Jing, X.; Ni, Q. Anti-Intelligent UAV Jamming Strategy via Deep Q-Networks. *IEEE Trans. Commun.* **2019**, *68*, 569–581. [CrossRef]
2. Kirk, B.H.; Narayanan, R.O.M.; Gallagher, K.A.; Martone, A.F.; Sherbondy, K.D. Avoidance of Time-Varying Radio Frequency Interference with Software-Defined Cognitive Radar. *IEEE Trans. Aerosp. Electron. Syst.* **2019**, *55*, 1090–1107. [CrossRef]
3. Jiang, X.F.; Zhou, F.; Jian, Y.; Xi, H.S. An Optimal POMDP-based Anti-Jamming Policy for Cognitive Radar. In Proceedings of the 13th IEEE Conference on Automation Science and Engineering (CASE), Xi'an, China, 20–23 August 2017; pp. 938–943.
4. Li, K.; Jiu, B.; Liu, H.W. Game Theoretic Strategies Design for Monostatic Radar and Jammer Based on Mutual Information. *IEEE Access* **2019**, *7*, 72257–72266. [CrossRef]
5. Zhang, X.X.; Ma, H.; Zhou, S.H.; Liu, H.W. Game Theory Design for Deceptive Jamming Suppression in Polarization MIMO Radar. *IEEE Access* **2019**, *7*, 114191–114202. [CrossRef]
6. Jiang, X.F.; Zhou, F.; Chen, S.W.; He, H.S.; Yang, J. Jamming Resilient Tracking Using POMDP-Based Detection of Hidden Targets. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 983–998. [CrossRef]
7. Olivier, S.; Olivier, B. *Markov Decision Processes in Artificial Intelligent*; ISTE Ltd.: London, UK; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010.
8. Yu, S.L.; Liu, Z.M.; Wu, J.B. Strategic behavior in the partially observable Markovian queues with partial breakdowns. *Oper. Res. Lett.* **2017**, *45*, 471–474. [CrossRef]
9. Mitola, J. Cognitive Radio. Ph.D. Thesis, Royal Institute of Technology, Stockholm, Sweden, 1999.
10. Chen, W.; Sun, H.Z.; Qi, E.Y.; Shen, K. Key Technology Prospects of Radar Seeker Signal Processing in Intelligent Age. *Areo Weapon.* **2019**, *26*, 76–82.
11. Liu, M.; Qi, H.Y.; Wang, J.N.; Liu, L.Z. Design of intelligence anti-jamming strategy based on neural network algorithm. *Comput. Meas. Control.* **2018**, *26*, 155–159.
12. Wu, X.M. Intelligent Anti-jamming Technology of Low-altitude Radar. *Ship Electron. Eng.* **2018**, *38*, 79–81.
13. Wang, X.; Qin, K.; Qin, Y.W. Intelligent Anti–Jamming System of Rader Based on Jamming Recognition. *Electron. Inf. Warf. Technol.* **2018**, *33*, 48–52.
14. He, B.; Sun, H.T. A Review of Game Theory Analysis in Cognitive Radar Anti-jamming. *J. Electron. Inf. Technol.* **2021**, *43*, 1199–1211.
15. Wang, H.; Wang, F. Application of Reinforcement Learning Algorithms in anti-jamming of Intelligent Radar. *Mod. Radar* **2020**, *42*, 40–44+48. [CrossRef]
16. Yuan, Q. Anti-Active Jamming Method of Intelligent Radar Network. Master's Thesis, Harbin Institute of Technology, Harbin, China, 2019.
17. Ak, S.; Bruggenwirth, S. Avoiding Jammers: A Reinforcement Learning Approach. In Proceedings of the 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 28–30 April 2020; pp. 321–326.
18. Shani, G.; Pineau, J.; Kaplow, R. A survey of point-based POMDP solvers. *Auton. Agents Multi-Agent Syst.* **2013**, *27*, 1–51. [CrossRef]
19. Liu, Y.; Zheng, J.; Chang, F. Learning and planning in partially observable environments without prior domain knowledge. *Int. J. Approx. Reason.* **2022**, *142*, 147–160. [CrossRef]
20. Smallwood, R.D.; Sondik, E.J. The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon. *Oper. Res.* **1973**, *21*, 1071–1088. [CrossRef]
21. Sondik, E.J. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Oper. Res.* **1978**, *26*, 282–304. [CrossRef]
22. Kurniawati, H.; Hsu, D.; Lee, W.S. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*; MIT Press: Cambridge, MA, USA, 2009; pp. 65–72.
23. Zhou, M.C. Research on Radar Jamming Decision Technology Based on Game Theory. Master's Thesis, Xidian University, Xi'an, China, 2014.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.