З.Б.СОХОВА[1], **Р.Р.ШИКЗАТОВ**[2]
[1]*НИИ Системных исследований РАН, Москва*
[2]*Национальный исследовательский ядерный университет «МИФИ», Москва*
zarema_s@mail.ru, ridvansmail@gmail.com

МОДЕЛЬ КООПЕРИРУЮЩИХСЯ АГЕНТОВ-ОХРАННИКОВ С ПОТРЕБНОСТЯМИ И МОТИВАЦИЯМИ

Построена и исследована компьютерная модель автономных агентов-охранников, функционирующих совместно в круговом кольце, разбитом на сектора. Проведены компьютерные эксперименты с тремя типами модели: агентов с мотивациями, агента без мотиваций и группы агентов с мотивациями. Показано, что кооперирующиеся агенты-охранники с мотивациями более успешно справляются со своей функцией – поиска нарушителей и поддержания внутренней энергии на нужном уровне.

Ключевые слова: модельные организмы, потребности, мотивации, обучение с подкреплением, охранное поведение.

1. Введение

В настоящей работе исследуется роль мотиваций и кооперации в поведении автономного агента-охранника, проводится сравнительный анализ модели с мотивациями и без мотиваций а так же проверяется предположение о том, что кооперация повышает эффективность охранного поведения. Работа развивает модели [1, 2], в которых было начато исследование агентов, обладающих естественными потребностями и мотивациями.

Считаем, что агент-охранник имеет две потребности: *питания* и *охраны территории*. Каждой потребности соответствует определенная мотивация. Имеются розетки, от которых агент может пополнить свой ресурс. И случайные нарушители, которых агент может прогонять.

Функция агента-охранника состоит в поиске нарушителей и поддержании внутренней энергии на нужном уровне. Задача агента – минимизация числа нарушителей.

2. Модель агента охранника с двумя потребностями

2.1. Модельный мир агента-охранника. Рассматривается один агент-охранник, обладающий внутренним ресурсом $R(t)$. Время t дискретно. Моделью мира агента является круговое кольцо, разбитое на шесть

секторов. Кольцо представляет собой границу охраняемой территории. В четных секторах находятся розетки. В каждый такт времени в любом из секторов с вероятностью p_1 появляется нарушитель.

Агент-охранник обладает мотивациями, соответствующими потребностям. В каждый такт времени одна из потребностей и соответствующая ей мотивация агента являются ведущими.

Потребностям агента соответствуют два фактора: фактор питания F_f и фактор охраны территории F_p. Фактор питания пропорционален ресурсу агента: $F_f = k_F R(t)$. Фактор охраны территории увеличивается при выполнении агентом действия «удар» и наличия нарушителя в одной клетке с агентом на ΔF_p и уменьшается на 1 в других ситуациях.

Удовлетворение ведущей потребности является положительным подкреплением при обучении.

2.2. Система управления агента-охранника с двумя потребностями. Система управления агента основана на наборе правил вида: $S_k \rightarrow A_k$, где S_k — ситуация, A_k — действие, k — номер правила. Согласно правилам в ситуации S_k нужно выполнить действие A_k. Каждое правило имеет свой вес W_k. Веса правил изначально случайны, а затем модифицируются методом обучения с подкреплением.

В каждый такт времени агент может выполнять одно из следующих действий A_k: 1) питание, 2 и 3) перемещение на один сектор по или против часовой стрелки соответственно, 4) удар, 5) отдых. Если действие «питание» вырабатывается в секторе, где имеется розетка, то ресурс агента $R(t)$ увеличивается на r'_1, иначе – уменьшается на r_1. При выполнении действий «перемещение» (в любом из двух направлений), удар и отдых ресурс агента уменьшается на величины r_2, r_3, r_4, r_5.

Ситуация S_k определяется 1) наличием или отсутствием розетки в текущем секторе и в двух соседних секторах, 2) наличием или отсутствием нарушителя в текущем секторе и в двух соседних секторах, и 3) ведущей мотивацией.

Если ресурс агента R меньше порога r_{th1}, то ведущей мотивацией является мотивация питания M_f, иначе ведущей является мотивация охраны M_p.

Каждый такт времени с вероятностью $1-\varepsilon$ выполняется то действие, для которого вес W_k соответствующего ему правила для текущей ситуации максимален, с вероятностью ε выполняется случайное действие.

2.3. Схема обучения. Используется схема обучения с подкреплением [1, 3]. Подкреплением является изменение фактора ведущей мотивации F_f или F_p:

$$\Delta W(t-1) = \alpha[F_L(t) - F_L(t-1) + \gamma W(t) - W(t-1) \quad (1)$$

где $F_L(t)$ — фактор ведущей в такт t мотивации, $W(t)$ и $W(t\text{-}1)$ — веса правил, примененных в такты t и t-1, α – параметр скорости обучения, γ — дисконтный фактор.

2.4. Результаты моделирования. Параметры компьютерного моделирования составляли: $\Delta F_p = 5$, $k_F = 0{,}2$, ε = 0.05, γ = 0,9, α = 0,1, $r_{th1} = 50.0$, $r'_1 = 50$, $r_1 = r_2 = r_3 = r_4 = 1$, $r_5 = 5$, $p_1 = 0{,}1$ либо $p_1 = 0{,}01$.

Результаты моделирования при $p_1 = 0{,}1$ представлены на рис. 1, 2.

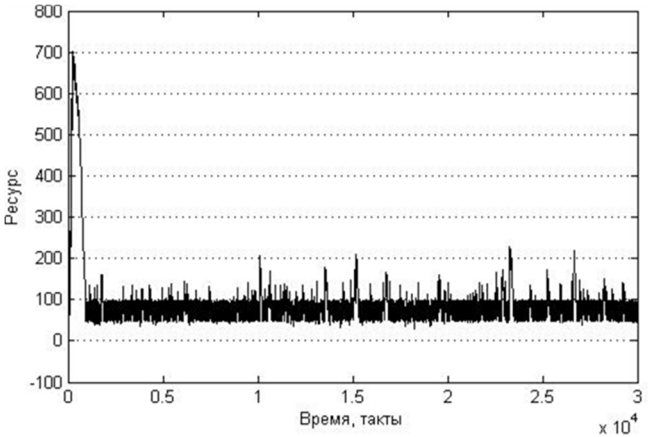

Рис. 1. Динамика ресурса агента-охранника с мотивациями.

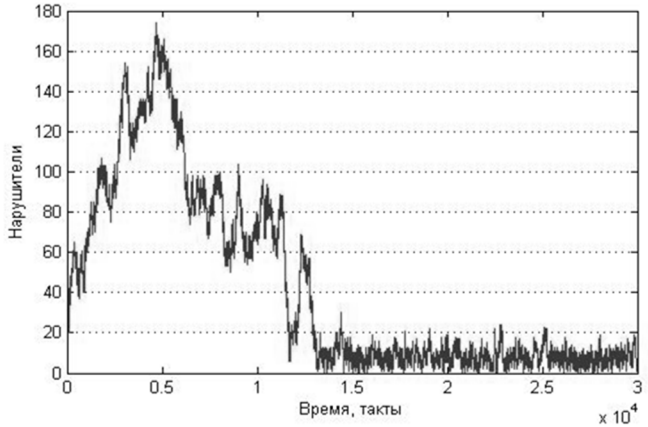

Рис. 2. Динамика количества нарушителей в модели с мотивациями.

Видно, что агент-охранник обучается поддерживать внутренний ресурс на уровне не ниже заданного порога. Количество нарушителей также уменьшается. При p_1 = 0,01 наблюдается аналогичная картина, только количество нарушителей для обученного агента становится значительно меньше.

При моделировании также наблюдалось, что обучение приводило к формированию различных независимых цепочек действий. Например, если ведущая мотивация охрана, и в текущем секторе есть нарушитель, система управления агента вырабатывает действие «удар». Если агент не видит нарушителя в текущем секторе, но нарушитель есть в одном из соседних секторов, то агент-охранник выбирает действие «двигаться» в сектор с нарушителем, а затем действие «удар». В ситуации, когда ресурс меньше либо равен порогу, агент выбирает действие питаться, несмотря на наличие нарушителей в текущем секторе.

3. Модель агента-охранника без мотиваций

3.1. Система управления агента-охранника без мотиваций. Система управления агента аналогична системе управления агента с мотивациями. Вектор ситуации S_k определяется 1) наличием или отсутствием розетки в текущем секторе и в двух соседних секторах, 2) наличием или отсутствием нарушителя в текущем секторе и в двух соседних секторах, и в отличие от модели, описанной выше, не содержит ведущую мотивацию.

При выполнении действия «питание» и наличии розетки в секторе ресурс агента увеличивается на dR_f, иначе – уменьшается на dR'_f. При выполнении действия «удар» и наличии нарушителя в секторе ресурс агента увеличивается на dF_s. При питании, перемещении в любом из двух направлений, ударе или отдыхе ресурс агента уменьшается соответственно на r_0, r_1, r_2, r_3, r_4.

3.2. Схема обучения. В данном варианте модели подкреплением является изменение ресурса $R(t)$:

$$\Delta W(t-1) = \alpha[R(t) - R(t-1) + \gamma W(t) - W(t-1) \qquad (2)$$

где $W(t)$ и $W(t\text{-}1)$ — веса правил, примененных в такты t и t-1, α — параметр скорости обучения, γ — дисконтный фактор.

3.3. Результаты моделирования. Параметры компьютерного моделирования составляли: $dR_f = 1$, $dR_s = 3$, $r_0=r_1=r_2=r_3=1$, ε = 0,05, γ = 0,9, α = 0,3, p_1 = 0,1, либо p_1 = 0,01. Результаты моделирования при p_1 = 0,1 представлены на рис. 3, 4.

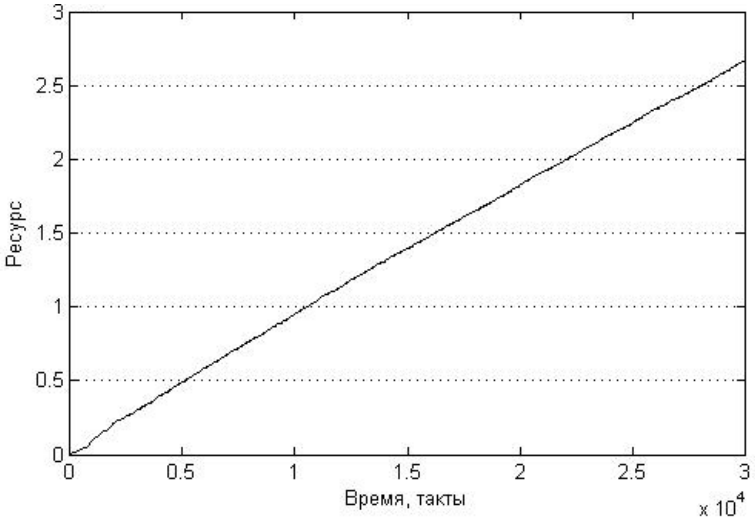

Рис. 3. Динамика ресурса агента-охранника без мотиваций.

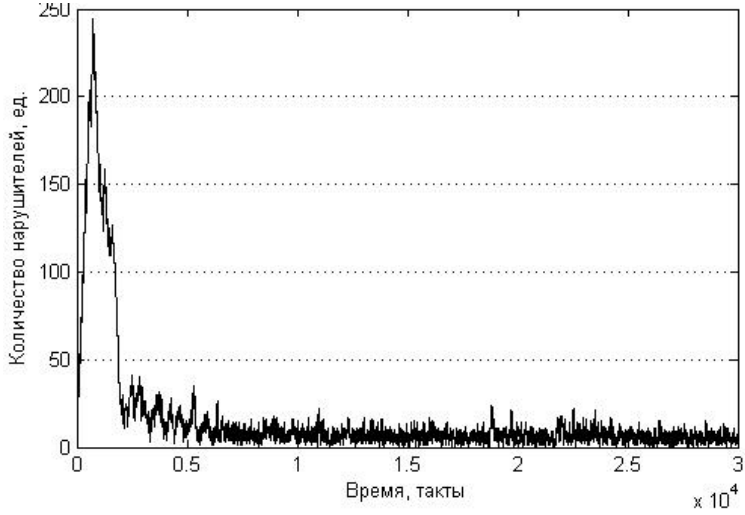

Рис. 4. Динамика количества нарушителей в модели без мотиваций.

Видно, что ресурс агента растет. Обученный агент справляется со своей задачей — количество агентов-нарушителей резко уменьшается.

4. Модель кооперирующихся агентов охранников.

4.1. Система управления кооперирующихся агентов-охранников.

Рассматриваются шесть взаимодействующих агентов-охранников. Система управления каждого из агентов аналогична системе управления агента с мотивациями. Изменена структура сенсоров: ситуация S_k определяется 1) наличием или отсутствием розетки в текущем секторе и в двух соседних секторах, 2) наличием или отсутствием нарушителя в текущем секторе и в двух соседних секторах, и 3) наличием или отсутствием других агентов в текущем секторе и в двух соседних секторах 4) ведущей мотивацией.

Наличие возможности агентам «видеть» других охранников рассматривается как условие возникновения кооперации.

Потребностям агента, аналогично модели анимата с мотивациями, соответствуют два фактора: фактор питания F_f и фактор охраны территории F_p. Фактор питания пропорционален ресурсу агента: $F_f = k_F R(t)$. Фактор охраны территории увеличивается при выполнении агентом действия «удар» и наличия нарушителя в одной клетке с агентом на ΔF_p и уменьшается на 1 в других ситуациях.

Удовлетворение ведущей потребности является положительным подкреплением при обучении. Схема обучения построена в соответствии с формулой 1.

4.2. Результаты моделирования.

Параметры компьютерного моделирования составляли: $\Delta F_p = 5$, $k_F = 0{,}2$, $\varepsilon = 0.05$, $\gamma = 0{,}9$, $\alpha = 0{,}1$, $r_{th1} = 50.0$, $r'_1 = 50$, $r_1 = r_2 = r_3 = r_4 = 1$, $r_5 = 5$. $p_1 = 0{,}1$, либо $p_1 = 0{,}01$. Результаты моделирования представлены на рис.5.

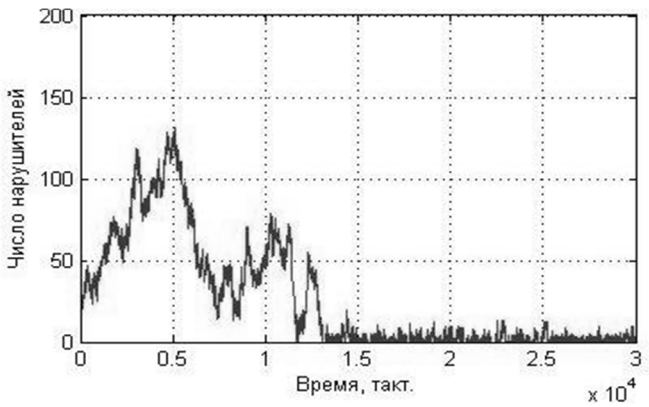

Рис. 5. Динамика количества нарушителей в модели с кооперацией.

Из рисунка, по моменту спада числа нарушителей видно, что обучение происходит на более продолжительных интервалах времени, это связано с тем, что число правил, которые необходимо обучить увеличилось почти на порядок. Мультимодальности графика, связанной с процессами обучения правил кооперации по каждому агенту, может не наблюдаться.

Результаты моделирования были сравнены с аналогичными (рис. 6) для модели с шестью агентами без возможности кооперации.

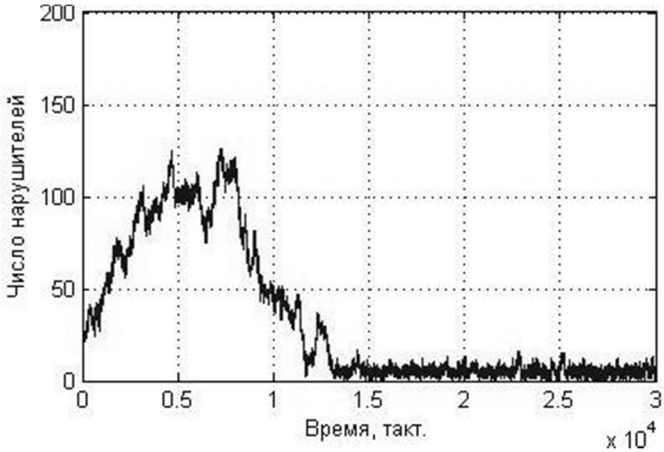

Рис. 6. Динамика количества нарушителей в модели без кооперации

Видно, что агенты без способности к кооперации быстрее обучаются.

Среднее значение числа нарушителей для участков графиков после 15000 тактов для модели с кооперацией несколько меньше, чем аналогичное для модели без кооперации.

5. Заключение

Построена и исследована модель автономных агентов-охранников с мотивациями и без мотиваций. Сравнительный анализ вариантов модели показал, что в модели без мотиваций количество нарушителей остающихся в мире незначительно меньше, чем в модели с мотивациями. В модели с мотивациями агент-охранник показывает более разумное поведение, так как не выполняет лишних действие «питание», а питается только, когда ресурс становится ниже порога. Дальнейшее развитие модели показало, что увеличение числа агентов увеличивает

эффективность охраны территории, а кооперация шести агентов охранников позволяет несколько более эффективно выполнять задачу.

Авторы благодарны В.Г.Редько за ряд полезных консультаций.

Список литературы

1. Редько В.Г., Бесхлебнова Г.А. Моделирование адаптивного поведения автономных агентов. — Нейрокомпьютеры: разработка, применение. 2010. № 3. С. 33–38

2. Коваль А. Г., Редько В. Г. Поведение модельных организмов, обладающих естественными потребностями и мотивациями. — «Математическая биология и биоинформатика», 2012. Т. 7. № 1. С. 266-273.

3. Sutton R., Barto A. Reinforcement Learning: An Introduction. — Cambridge: MIT Press, 1998.