МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА, 2007, том 2, №1, с.160-180,
http://www.matbio.org/downloads/Redko2007(2_160).pdf

==========МАТЕРИАЛЫ I МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ==========
==========«МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА»==========

УДК: 004.81:004.032.26

Модели адаптивного поведения и проблема происхождения интеллекта*

В.Г. Редько**

Научно-исследовательский институт системных исследований, Российская академия наук, Москва, 119333, Россия

Аннотация. Обсуждается подход к исследованию проблемы происхождения интеллекта на основе построения моделей эволюции адаптивного поведения. Характеризуются работы ведущих лабораторий в области моделирования адаптивного поведения. Особое внимание уделяется методу обучения с подкреплением, в частности, нейросетевым адаптивным критикам. Излагается проект «Мозг анимата», нацеленный на формирование общей «платформы» для систематического построения моделей адаптивного поведения. Приводятся результаты исследования конкретной модели эволюции самообучающихся агентов на основе нейросетевых адаптивных критиков. В порядке обсуждения предлагается программа будущих исследований эволюции адаптивного поведения.

Ключевые слова: *адаптивное поведение, биологические информационные системы, проблема происхождения интеллекта, когнитивная эволюция, обучение с подкреплением*

1. ВВЕДЕНИЕ. БИОЛОГИЯ И ИНФОРМАТИКА – НАУКИ 21-ГО ВЕКА. ЧТО НА СТЫКЕ?

Настоящая статья посвящена анализу проблем, которые связаны информационными процессами в биологических организмах, с тем, как информационные процессы обеспечивают приспособление живых организмов к переменной внешней среде.

На стыке биологии и информатики есть ряд глубоких проблем, имеющих мировоззренческое значение, исследованию которых до сих пор уделяется несоизмеримо малое внимание по сравнению с их значимостью.

Действительно, в процессе биологической эволюции возникли чрезвычайно сложные и вместе с тем удивительно эффективно функционирующие живые организмы. Эффективность, гармоничность и согласованность работы «компонент» живых существ обеспечивается биологическими управляющими системами. Но каковы эти управляющие системы? Как и почему они эволюционно возникли? Какие информационные процессы обеспечивают работу этих управляющих систем? Как животные познают внешний мир и используют это познание для управления своим поведением? Как эволюционное развитие познавательных способностей животных привело к возникновению интеллекта человека? До какой степени исследования причин возникновения естественного интеллекта могут способствовать развитию искусственного интеллекта?

---

* Работа выполнена при финансовой поддержке программы Президиума РАН «Интеллектуальные компьютерные системы» (проект 2-45) и РФФИ (проект № 07-01-00180).

** vgredko@gmail.com

Среди спектра подобных вопросов особое место занимают две крупные проблемы, исследование которых важно с точки зрения развития научного миропонимания:

- проблема происхождения молекулярно-генетических систем управления живыми клетками в процессе происхождения жизни,
- проблема происхождения интеллекта человека.

В настоящей статье обсуждаются подходы к анализу второй из этих проблем на основе моделей адаптивного поведения. Относительно первой проблемы здесь только отметим, что есть ряд математических моделей (модели квазивидов [1-3], гиперциклов [2], сайзеров [4]), характеризующих постепенное усложнение информационно-кибернетических систем и позволяющих представить гипотетические этапы предбиологической эволюции (см. обзор этих моделей в [5, 6]).

Структура статьи следующая. В разделе 2 рассматривается гносеологическая проблема, подчеркивающая важность исследования проблемы происхождения интеллекта человека путем моделирования когнитивной эволюции. Раздел 3 содержит краткий обзор направления исследований «Адаптивное поведение», которое может рассматриваться как задел разработок моделей когнитивной эволюции. Основной подход этого направления – моделирование искусственных «организмов», способных приспосабливаться к внешней среде. Эти организмы часто называются «*аниматами*» (от англ. animal + robot = animat). В разделе 4 излагается один из основных методов обучения аниматов – метод обучения с подкреплением, в частности, излагаются схемы и принципы функционирования нейросетевых адаптивных критиков (схемы обучения с подкреплением на основе нейронных сетей). Раздел 5 описывает версию проекта «Мозг анимата», использующую нейросетевые адаптивные критики. Этот проект нацелен на формирование общей «платформы» для систематического построения моделей адаптивного поведения. Приводятся результаты исследования конкретной модели эволюции самообучающихся адаптивных агентов на основе нейросетевых адаптивных критиков. В разделе 6 в порядке обсуждения представлены контуры программы будущих исследований проблемы происхождения интеллекта.

2. МОЖНО ЛИ ОБОСНОВАТЬ МАТЕМАТИЧЕСКУЮ СТРОГОСТЬ?

Каждый, кто достаточно серьезно изучал классический математический анализ, мог по достоинству оценить красоту математической строгости. Благодаря работам О. Коши, Б. Больцано, К. Вейерштрасса и других математиков XIX века, одна из наиболее содержательных частей математики – дифференциальное и интегральное исчисление – получила столь серьезное обоснование, что невольно возникает желание распространить подобную строгость на возможно большую часть человеческих знаний. Однако если посмотреть широко на естественные науки в целом, то может возникнуть вопрос: а насколько вообще обоснована применимость математики к познанию природы? Ведь те процессы, которые происходят в мышлении математика, совсем не похожи на те процессы, которые происходят в природе и изучаются естествоиспытателями.

Действительно, рассмотрим физику, наиболее фундаментальную из естественнонаучных дисциплин. Мощь физики связана с эффективным применением математики. Но математик строит свои теории чисто логическим путем, совсем независимо от внешнего мира, используя свое мышление (в тиши кабинета, лежа на диване, в изолированной камере…). Почему же результаты, получаемые математиком, применимы к реальной природе?

Итак, возникает определенное сомнение в обоснованности самой математической строгости. В более общей формулировке проблему можно поставить так: *почему логика человеческого мышления применима к познанию природы?* Действительно, с одной стороны, логические процессы вывода происходят в нашем, человеческом мышлении, с

другой стороны, процессы, которые мы познаем посредством логики, относятся к изучаемой нами природе. Эти два типа процессов различны. Поэтому далеко не очевидно, что мы можем использовать процессы первого типа для познания процессов второго типа.

Можно ли конструктивно подойти к решению этих вопросов? Скорее всего, да. Чтобы продемонстрировать такую возможность, будем рассуждать следующим образом.

Рассмотрим одно из элементарных правил, которое использует математик в логических выводах, правило modus ponens: «если имеет место *A*, и из *A* следует *B*, то имеет место *B*», или {*A*, *A* --> *B*} => *B*.

А теперь перейдем от математика к собаке И.П. Павлова. Пусть у собаки вырабатывают условный рефлекс, в результате в памяти собаки формируется связь «за УС должен последовать БС» (УС - условный стимул, БС - безусловный стимул). И когда после выработки рефлекса собаке предъявляют УС, то она, «помня» о хранящейся в ее памяти «записи»: УС --> БС, делает элементарный «вывод» {УС, УС --> БС } => БС. И у собаки, ожидающей БС (скажем, кусок мяса), начинают течь слюнки.

Конечно, применение правила modus ponens (чисто дедуктивное) математиком и индуктивный «вывод», который делает собака, явно различаются. Но можем мы ли думать об эволюционных корнях логических правил, используемых в математике? Да, вполне можем – умозаключение математика и «индуктивный вывод» собаки качественно аналогичны.

Итак, мы можем думать над эволюционными корнями логики, мышления, интеллекта. И более того, было бы очень интересно попытаться строить модели эволюционного происхождения мышления. По-видимому, наиболее четкий путь такого исследования – построение математических и компьютерных моделей «интеллектуальных изобретений» биологической эволюции, таких как безусловный рефлекс, привыкание, классический условный рефлекс, инструментальный условный рефлекс, цепи рефлексов, …, логика [7]. То есть, целесообразно с помощью моделей представить общую картину когнитивной эволюции – эволюции когнитивных способностей животных и эволюционного происхождения интеллекта человека.

Есть ли задел таких исследований? Оказывается, что да, есть. Сравнительно недавно сформировалось направление исследований «Адаптивное поведение», дальняя цель которого очень близка к задаче моделирования когнитивной эволюции.

3. НАПРАВЛЕНИЕ ИССЛЕДОВАНИЙ «АДАПТИВНОЕ ПОВЕДЕНИЕ»

С начала 1990-х годов за рубежом активно развивается направление исследований «Адаптивное поведение» (АП) [8-10]. Основной подход этого направления – конструирование и исследование искусственных (в виде компьютерной программы или робота) «организмов», способных приспосабливаться к внешней среде. Эти организмы называются «*аниматами*» (от англ. animal + robot = animat) или «агентами».

Поведение аниматов имитирует поведение животных. Исследователи направления АП стараются строить именно такие модели, которые применимы к описанию поведения *как реального животного, так и искусственного анимата* [11, 12].

Программа-минимум направления «Адаптивное поведение» – исследовать архитектуры и принципы функционирования, которые позволяют животным или роботам жить и действовать в переменной внешней среде.

Программа-максимум этого направления – попытаться проанализировать эволюцию когнитивных способностей животных и эволюционное происхождение человеческого интеллекта [13].

Программа-максимум близка к очерченной выше задаче моделирования когнитивной эволюции.

Для исследований АП характерен *синтетический подход*: здесь конструируются архитектуры, обеспечивающие «интеллектуальное» поведение аниматов. Причем это конструирование проводится как бы с точки зрения инженера: исследователь сам «изобретает» архитектуры, подразумевая конечно, что какие-то подобные структуры, обеспечивающие адаптивное поведение, должны быть у реальных животных. При этом направление исследований АП рассматривается как бионический подход к разработке систем искусственного интеллекта [14]. Хотя «официально» направление АП было провозглашено в 1990 году, были явные провозвестники этого направления. Приведем примеры из истории отечественной науки.

В 1960-х годах блестящий кибернетик и математик М.Л. Цетлин предложил и исследовал модели автоматов, способных адаптивно приспосабливаться к окружающей среде. Работы М.Л. Цетлина инициировали целое научное направление, получившее название «коллективное поведение автоматов» [15, 16]. В 1960-70-х годах под руководством талантливого кибернетика М.М. Бонгарда был предложен интересный проект «Животное», направленный на моделирование адаптивного поведения искусственных организмов с иерархией целей и подцелей [17, 18]. Хороший обзор ранних работ по адаптивному поведению, представлен в книге М.Г. Гаазе-Рапопорта, Д.А. Поспелова «От амебы до робота: модели поведения» [18].

Подчеркнем, что АП – активно развивающееся направление исследований. Есть научное общество «The International Society for Adaptive Behavior» (http://www.isab.org.uk). Регулярно проводятся международные конференции «Simulation of Adaptive Behavior (From Animal to Animat)». Издается журнал «Adaptive Behavior» (http://www.isab.org.uk/journal).

В исследованиях АП используется ряд нетривиальных компьютерных методов:
- нейронные сети,
- генетический алгоритм и другие методы эволюционной оптимизации [19-21],
- классифицирующие системы (Classifier Systems) [22],
- обучение с подкреплением (Reinforcement Learning) [23].

Подчеркнем, что в АП в основном используется *феноменологический подход* к исследованиям систем управления адаптивным поведением. Предполагается, что существуют формальные правила адаптивного поведения, и эти правила не обязательно связаны с конкретными микроскопическими нейронными или молекулярными структурами, которые есть у живых организмов. Скорее всего, такой феноменологический подход для исследований АП вполне имеет право на существование. В пользу этого тезиса приведем аналогию из физики. Есть термодинамика, и есть статистическая физика. Термодинамика описывает явления на феноменологическом уровне, статистическая физика характеризует те же явления на микроскопическом уровне. В физике термодинамическое и стат-физическое описания относительно независимы друг от друга и, вместе с тем, взаимодополнительны. По-видимому, и для описания живых организмов может быть аналогичное соотношение между феноменологическим (на уровне поведения) и микроскопическим (на уровне нейронов и молекул) подходами. При этом, естественно ожидать, что для исследования систем управления адаптивным поведением феноменологический подход на начальных этапах работ должен быть более эффективен, так как очень трудно сформировать целостную картину поведения на основе анализа всего сложного многообразия функционирования нейронов, синапсов, молекул.

В настоящее время исследования адаптивного поведения включают в себя работы по следующим темам [24]:
- сенсорные системы и управление,
- обучение и адаптация,
- выбор действий, навигация и внутренние модели мира,

- антиципаторное адаптивное поведение,
- нейроэволюция (настройка нейронных сетей аниматов эволюционными методами),
- возникновение языка и коммуникаций при адаптивном поведении,
- коллективное и социальное поведение,
- биологически инспирированное адаптивное поведение роботов,
- поведение и мышление как сложные адаптивные системы.

Исследования по адаптивному поведению ведутся в ряде университетов и лабораторий, таких как:

- AnimatLab (Париж, руководитель – один из инициаторов данного направления Жан-Аркадий Мейер) [8, 13, 25].
- Лаборатория искусственного интеллекта в университете Цюриха (руководитель Рольф Пфейфер) [26, 27].
- Лаборатория искусственной жизни и роботики в Институте когнитивных наук и технологий (Рим, руководитель Стефано Нолфи) [28, 29], ведущая исследования в области эволюционной роботики и принципов формирования адаптивного поведения.
- Лаборатория информатики и искусственного интеллекта в Массачусетском технологическом институте (руководитель Родни Брукс) [30, 31], которая ведет исследования широкого спектра интеллектуальных и адаптивных систем, включая создание интеллектуальных роботов.
- Институт нейронаук Дж. Эдельмана, где ведутся разработки поколений моделей работы мозга (Darwin I, Darwin II, …) и исследования поведения искусственного организма NOMAD (Neurally Organized Mobile Adaptive Device), построенного на базе этих моделей [32-35].

В России исследования адаптивного поведения пока ведутся скромными усилиями ученых-энтузиастов, среди этих работ следует отметить:

- модели поискового адаптивного поведения [11, 12, 36, 37] (В.А. Непомнящих, Институт биологии внутренних вод им. И.Д. Папанина РАН);
- концепции и модели автономного адаптивного управления на основе аппарата эмоций [38, 39] (А.А. Жданов, Институт системного программирования РАН);
- разработку принципов построения систем управления антропоморфных и гуманоидных роботов [40] (Л.А. Станкевич, Санкт-Петербургский политехнический университет);
- разработку нейросетевых моделей поведения роботов и робототехнических устройств [41] (А.И. Самарин, НИИ нейрокибернетики им. А.Б. Когана РГУ);
- модели адаптивного поведения на основе эволюционных и нейросетевых методов [42-46] (В.Г. Редько, М.С. Бурцев, О.П. Мосалов, НИИ системных исследований РАН, Институт прикладной математики им. М.В. Келдыша РАН).

Работы отечественных исследователей адаптивного поведения представлены в коллективной монографии [47].

Один из ключевых методов, используемых при разработке моделей адаптивного поведения – метод обучения с подкреплением. В следующем разделе мы охарактеризуем это метод, а также кратко опишем интересное направление работ, развиваемое в рамках теории обучения с подкреплением, – нейросетевые адаптивные критики.

4. ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

4.1. Общая схема обучения с подкреплением

Теория обучения с подкреплением (reinforcement learning) была развита в цикле работ Р. Саттона и Э. Барто (Массачусетский университет), который подробно отражен в книге [23].

Общая схема обучения с подкреплением показана на рис. 1. Рассматривается анимат, взаимодействующий с внешней средой. Время предполагается дискретным: t = 1,2,… В текущей ситуации $S(t)$ анимат выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $S(t+1)$. Подкрепление может быть положительным (награда), $r(t) > 0$, или отрицательным (наказание), $r(t) < 0$.

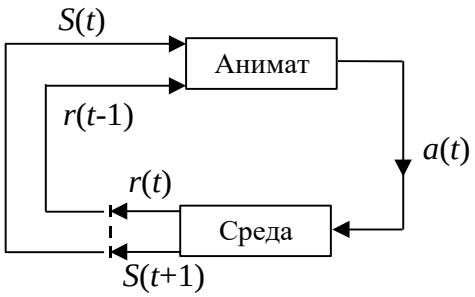

Рис. 1. Схема обучения с подкреплением.

Цель анимата – максимизировать суммарную награду, которую можно получить в будущем в течение длительного периода времени. Подразумевается, что анимат может иметь свою внутреннюю «субъективную» оценку суммарной награды и в процессе обучения постоянно совершенствует эту оценку. Эта оценка определяется с учетом дисконтного фактора:

$$U(t) = \sum_{j=0}^{\infty} \gamma^j r(t+j), \qquad t = 1,2,..., \tag{1}$$

где $U(t)$ – оценка суммарной награды, ожидаемой после момента времени t, γ – дисконтный фактор, $0 < \gamma < 1$. Дисконтный фактор учитывает, что чем дальше анимат «заглядывает» в будущее, тем меньше у него уверенность в оценке награды («рубль сегодня стоит больше, чем рубль завтра»).

Если множество возможных ситуаций $\{S_i\}$ и действий $\{a_j\}$ конечно, то существует простой метод обучения SARSA, каждый шаг которого соответствует цепочке событий $S(t) \rightarrow a(t) \rightarrow r(t) \rightarrow S(t+1) \rightarrow a(t+1)$.

4.2. Метод SARSA

Кратко опишем метод SARSA. В этом методе итеративно формируются оценки величины суммарной награды $Q(S(t), a(t))$, которую получит анимат, если в ситуации $S(t)$ он выполнит действие $a(t)$. Математическое ожидание суммарной награды равно:

$$Q(S(t), a(t)) = \mathrm{E}\{r(t) + \gamma\, r(t+1) + \gamma^2\, r(t+2) + \ldots\} \mid S = S(t), a = a(t). \tag{2}$$

Из (1) и (2) следует $Q(S(t), a(t)) = E\,[r(t) + \gamma\, Q(S(t+1), a(t+1))]$. Ошибку естественно определить так [23]:

$$\delta(t) = r(t) + \gamma\, Q(S(t+1), a(t+1)) - Q(S(t), a(t)). \tag{3}$$

Величина $\delta(t)$ называется ошибкой временной разности.

Здесь $\delta(t)$ – разность между той оценкой суммарной величины награды, которая формируется у анимата для момента времени t после выбора действия $a(t+1)$ в следующей ситуации $S(t+1)$ в момент времени $t+1$, и предыдущей оценкой этой же

величины, которая была у анимата в момент времени t. Предыдущая оценка равна $Q(S(t), a(t))$, новая оценка равна $r(t) + \gamma\, Q(S(t+1), a(t+1))$, что и отражает формула (3) для величины $\delta(t)$. В соответствии с этим $\delta(t)$ анимат и обучается (см. ниже, формулу (4)).

Каждый такт времени происходит как выбор действия, так и обучение анимата. Выбор действия происходит так:

– в момент t с вероятностью $1 - \varepsilon$ выбирается действие, соответствующее максимальному значению $Q(S(t), a_i)$: $a(t) = a_k$, $k = \arg\max_i \{Q(S(t), a_i)\}$
– с вероятностью ε выбирается произвольное действие случайным образом, $0 < \varepsilon << 1$. Такую схему выбора действия называют «ε–жадным правилом».

Обучение, т.е. переоценка величин $Q(S, a)$ происходит в соответствии с оценкой ошибки $\delta(t)$ – к величине $Q(S(t), a(t))$ добавляется величина, пропорциональная ошибке временной разности $\delta(t)$:

$$\Delta Q(S(t), a(t)) = \alpha\, \delta(t) = \alpha\, [r(t) + \gamma\, Q(S(t+1), a(t+1)) - Q(S(t), a(t))], \qquad (4)$$

где α – параметр скорости обучения.

Так как число ситуаций и действий конечно, то здесь происходит формирование матрицы $Q(S_j, a_i)$, соответствующей всем возможным ситуациям S_j и всем возможным действиям a_i. Оценка суммарной награды $Q(S(t), a(t))$ может рассматриваться как оценка качества действия $a(t)$ в текущей ситуации $S(t)$.

Метод обучения с подкреплением идейно связан с методом динамического программирования, и в том и в другом случае общая оптимизация многошагового процесса принятия решения происходит путем упорядоченной процедуры одношаговых оптимизирующих итераций, причем оценки эффективности тех или иных решений, соответствующие предыдущим шагам процесса, переоцениваются с учетом знаний о возможных будущих шагах. Например, при решении задачи поиска оптимального маршрута в лабиринте от стартовой точки к определенной целевой точке сначала находится конечный участок маршрута, непосредственно приводящий к цели, а затем ищутся пути, приводящие к конечному участку, и т.д. В результате постепенно прокладывается трасса маршрута от его конца к началу. Обучение с подкреплением, адаптивные критики и подобные методы часто называют приближенным динамическим программированием [48].

Важное достоинство метода обучения с подкреплением – его простота. Анимат получает от учителя или из внешней среды только сигналы подкрепления $r(t)$. Здесь учитель поступает с обучаемым объектом примитивно: «бьет кнутом» (если действия объекта ему не нравятся, $r(t) < 0$), либо «дает пряник» (в противоположном случае, $r(t) > 0$), не объясняя обучаемому объекту, как именно нужно действовать. Это радикально отличает этот метод от таких традиционных в теории нейронных сетей методов обучения, как метод обратного распространения ошибок [49], для которого учитель точно определяет, что должно быть на выходе нейронной сети при заданном входе.

Метод обучения с подкреплением был исследован рядом авторов (см. подробную библиографию в [23]) и был использован многочисленных приложениях. Подчеркнем, что метод обучения с подкреплением может рассматриваться как развитие автоматной теории адаптивного поведения, разработанной в работах М.Л. Цетлина и его последователей [15, 16].

В свою очередь, метод обучения с подкреплением получил свое развитие в работах по адаптивным критикам, в которых рассматриваются методы обучения, использующие нейросетевые аппроксиматоры функций оценки качества функционирования анимата. Простейшие схемы адаптивных критиков рассмотрим в следующем разделе.

4.3. Нейросетевые адаптивные критики [50]

Конструкции адаптивных критиков можно рассматривать как схемы обучения с подкреплением в случае, когда ситуации и/или действия задаются векторами **S** и **A** и изложенная выше схема итеративного формирования матрицы $Q(S_j, a_i)$ не работает. При этом характеристики системы управления целесообразно представить с помощью параметрически задаваемых аппроксимирующих функций (например, с помощью искусственных нейронных сетей), а обучение проводить путем итеративной оптимизации параметров.

В конструкции систем управления аниматов на основе адаптивных критиков входят два важных блока: Критик и Контроллер (иногда используют также термин Актор).

Критик – блок системы управления, оценивающий качество ее работы.

Контроллер – блок системы управления, формирующий действия этой системы.

Ниже мы опишем две простые конструкции адаптивных критиков: Q-критик и V-критик. Обе конструкции используют нейросетевую аппроксимацию характеристик системы управления.

Q-критик. Схема Q-критика представлена на рис. 2. Предполагаем, что как Критик, так и Контроллер представляют собой нейронные сети, а именно, многослойные персептроны (такие же, какие используются в методе обратного распространения ошибки [49]) с весами синапсов $\mathbf{W_C}$ и $\mathbf{W_A}$, соответственно.

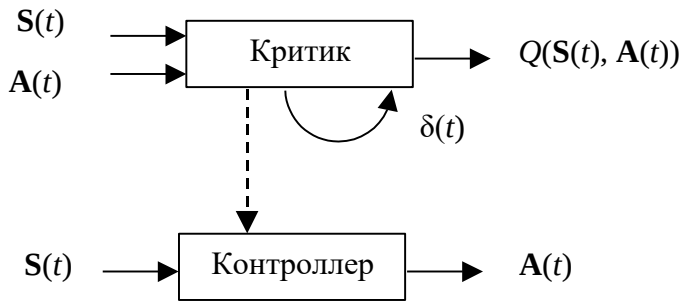

Рис. 2. Схема Q-критика. Вертикальная штриховая стрелка иллюстрирует то, что при обучении производные $\partial Q(t)/\partial A_k(t)$, вычисляемые в нейронной сети Критика, используются при настройке весов Контроллера (см. формулу (7)).

Функционирование этой схемы происходит следующим образом. В момент времени t Контроллер по вектору входной ситуации $\mathbf{S}(t)$ определяет вектор действия $\mathbf{A}(t)$ (команды на эффекторы). На входы Критика подаются два вектора: $\mathbf{S}(t)$ и $\mathbf{A}(t)$. По этому составному входному вектору Критик делает оценку качества $Q(t) = Q(\mathbf{S}(t),\mathbf{A}(t))$ действия $\mathbf{A}(t)$ в текущей ситуации $\mathbf{S}(t)$. Действие $\mathbf{A}(t)$ выполняется, анимат получает награду $r(t)$. Далее происходит переход к следующему моменту времени t+1. Все операции повторяются, в том числе делается оценка значения $Q(t+1)$. После этого для момента t определяется ошибка временной разности:

$$\delta(t) = r(t) + \gamma\, Q\,(t+1) - Q\,(t). \qquad (5)$$

Обучение нейросетей выполняется следующим образом:

$$\Delta\, \mathbf{W_C} = \alpha_1\, \mathrm{grad}_{WC}(Q\,(t))\, \delta(t), \qquad (6)$$

$$\Delta\, \mathbf{W_A} = \alpha_2\, \Sigma_k\, \{[\partial Q\,(t)\,/\partial\, A_k(t)]\, \mathrm{grad}_{WA}\, A_k(t)\,\}, \qquad (7)$$

где α_1 и α_2 – параметры скорости обучения. Производные по весам синапсов $\mathrm{grad}_{WC}(.)$ и $\mathrm{grad}_{WA}(.)$ в (6) и (7), а также $\partial Q(t)/\partial A_k(t)$ в (7) рассчитываются как производные сложных функций, аналогично тому, как это делается в методе обратного распространения ошибки [49]. В формуле (7) учитывается, что нужно брать

производные по всем компонентам вектора $\mathbf{A}(t)$ и суммировать по всем этим компонентам.

Смысл изменений весов синапсов по формулам (6), (7) состоит в том, что веса Критика и Контроллера меняются таким образом, чтобы уменьшить ошибку в оценке ожидаемой суммарной награды (обучение Критика) и увеличить значение самой награды при попадании анимата в близкие ситуации (обучение Контроллера).

V-критик. Схема V-критика представлена на рис. 3.

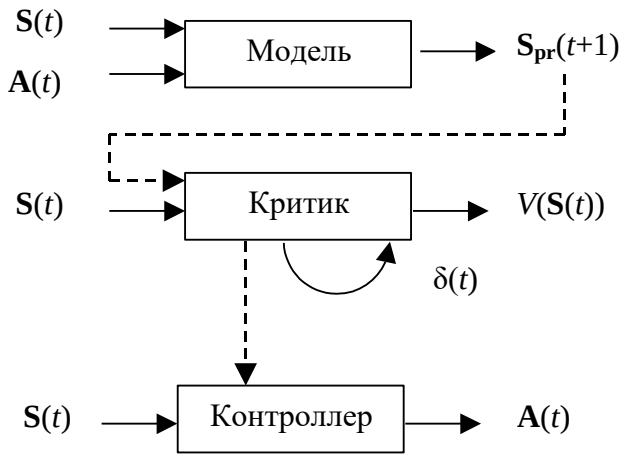

Рис. 3. Схема V-критика.

В этой схеме блок Критик, в отличие от схемы Q-критика, оценивает качество ситуации $V(\mathbf{S}(t))$ независимо от выполняемого действия. Однако эта схема содержит блок Модель, в котором прогнозируется будущая ситуация $\mathbf{S_{pr}}(t+1) = \mathbf{S_{pr}}(\mathbf{S}(t), \mathbf{A}(t))$ в зависимости от текущей ситуации $\mathbf{S}(t)$ и выполняемого действия $\mathbf{A}(t)$. И для этой прогнозируемой ситуации $\mathbf{S_{pr}}(t+1)$ блок Критик может сделать оценку ее качества $V_{\text{pr}} = V(\mathbf{S_{pr}}(t+1)) = V(\mathbf{S_{pr}}(\mathbf{S}(t), \mathbf{A}(t)))$, что аналогично оценке $Q(\mathbf{S}(t),\mathbf{A}(t))$, которую делает Q-критик. Предполагаем, что Критик, Контроллер и Модель представляют собой многослойные персептроны с весами синапсов $\mathbf{W_C}$, $\mathbf{W_A}$ и $\mathbf{W_M}$, соответственно.

Функционирование этой схемы происходит следующим образом. В момент времени t Контроллер по вектору входной ситуации $\mathbf{S}(t)$ определяет вектор действия $\mathbf{A}(t)$. Критик делает оценку качества $V(t) = V(\mathbf{S}(t))$ текущей ситуации $\mathbf{S}(t)$. Модель прогнозирует следующую ситуацию $\mathbf{S_{pr}}(t+1) = \mathbf{S_{pr}}(\mathbf{S}(t), \mathbf{A}(t))$. Критик оценивает качество прогнозируемой ситуации $V_{\text{pr}} = V(\mathbf{S_{pr}}(t+1))$. Действие $\mathbf{A}(t)$ выполняется, анимат получает награду $r(t)$. Оценивается ошибка временной разности:

$$\delta(t) = r(t) + \gamma\, V(\mathbf{S_{pr}}(t+1)) - V(\mathbf{S}(t)). \tag{8}$$

Обучаются Критик:

$$\Delta\, \mathbf{W_C} = \alpha_1\, \text{grad}_{\text{WC}}(V(t))\, \delta(t), \tag{9}$$

и Контроллер:

$$\Delta\, \mathbf{W_A} = \alpha_2\, \Sigma_k\, \{[\partial\, V(\mathbf{S_{pr}}(t+1))/\, \partial\, A_k(t)]\, \text{grad}_{\text{WA}}\, A_k(t)\}, \tag{10}$$

$$\partial\, V(\mathbf{S_{pr}}(t+1))/\, \partial\, A_k(t) = \Sigma_j\, \{[\partial\, V/\, \partial S_{pr\,j}\,]\, [\partial S_{pr\,j}\, /\partial\, A_k(t)]\}. \tag{11}$$

Производные в (11) берутся в соответствии с формулами нейронных сетей Критика и Модели.

Далее происходит переход к следующему моменту времени $t+1$. Сравниваются прогнозируемая $\mathbf{S_{pr}}(t+1)$ и реальная ситуация $\mathbf{S}(t+1)$. В соответствии с ошибкой этого прогноза обучается Модель обычным методом обратного распространения ошибки.

Обучение Критика состоит в том, чтобы итеративно уточнять оценку качества ситуаций $V(\mathbf{S}(t))$ в соответствии с поступающими подкреплениями. Обучение Контроллера состоит в том, чтобы постепенно формировать действия, приводящие к ситуациям с высокими значениями качества $V(\mathbf{S})$. Смысл обучения Модели – уточнение прогнозов будущих ситуаций. Отметим, что оценка функции качества ситуации $V(\mathbf{S}(t))$ в этой схеме аналогична эмоциональной оценке текущего состояния системы в моделях А.А. Жданова [38, 39]. Более полно теория адаптивных критиков характеризуется в работах [51, 52].

В следующем разделе излагается иерархическая архитектура системы управления аниматом на базе нейросетевых адаптивных критиков. Архитектура разрабатывается в рамках проекта «Мозг анимата» [53, 54], который основывается на теории функциональных систем П.К. Анохина.

5. ПРОЕКТ «МОЗГ АНИМАТА»[1]

Теория функциональных систем была предложена и развита в 1930-70 годах известным советским нейрофизиологом П.К. Анохиным [55]. Функциональная система по П.К. Анохину – схема управления, нацеленная на достижение полезных для организма результатов. Далее излагается основанный на теории функциональных систем проект «Мозг анимата».

5.1. Общая архитектура «Мозга анимата»

Предполагается, что система управления аниматом имеет иерархическую архитектуру. Базовым элементом системы управления является отдельная функциональная система (ФС).

Первый уровень (ФС1, ФС2, …) соответствует основным потребностям организма: питания, размножения, безопасности, накопления знаний. Более низкие уровни системы управления соответствуют тактическим целям поведения. Блоки всех этих уровней (включая первый) реализуются с помощью функциональных систем. Управление с верхних уровней может передаваться на нижние уровни (от «суперсистем» к «субсистемам») и возвращаться назад.

Предполагается, что система управления аниматом функционирует в дискретном времени; каждый такт времени активна только одна ФС.

Рассматривается простая формализация функциональной системы на основе нейросетевых адаптивных критиков. Функциональная система моделирует следующие важные особенности ее биологического прототипа: 1) прогноз результата действия, 2) сравнение прогноза и результата, 3) коррекцию прогноза путем обучения в соответствующих нейронных сетях, 4) принятие решения. Принятие решения в данной схеме ФС соответствует выбору одного из альтернативных действий.

Для развития проекта важно оценить возможности адаптивных критиков и проверить, как функционируют простые схемы адаптивных критиков в конкретных моделях. В следующем разделе излагаются результаты исследования такой модели.

5.2. Модель эволюции популяции самообучающихся агентов на базе нейросетевых адаптивных критиков [46, 56]

5.2.1. Описание модели

Исследуется модель эволюции популяции самообучающихся автономных агентов и анализируется взаимодействие между обучением и эволюцией. Модель отрабатывается на примере агента-брокера. Этот пример используется только для определенности, совершенно аналогично можно рассматривать функционирование модельного

[1] Термин «Мозг анимата» был предложен К.В. Анохиным

«организма», более подобного биологическим прототипам, например, «организма», помещенного во внешнюю среду, которая определяется зависимостью температуры от времени, аналогичной курсу акций для агента-брокера.

Схема агента. Рассматривается модель агента-брокера, который имеет ресурсы двух типов: деньги и акции; сумма этих ресурсов составляет капитал агента $C(t)$; доля акций в капитале равна $u(t)$. Внешняя среда определяется временным рядом $X(t)$; $t = 0,1,2,...$; $X(t)$ – курс акций на бирже в момент времени t . Агент стремится увеличить свой капитал $C(t)$, изменяя значение $u(t)$. Динамика капитала определяется выражением [57]:

$$C(t+1) = C(t)\,[1 + u(t+1)\,\Delta X(t+1) / X(t)], \qquad (12)$$

где $\Delta X(t+1) = X(t+1) - X(t)$ – текущее изменение курса акций. Используется логарифмическая шкала для ресурса агента, $R(t) = \log C(t)$ [58]. Текущее подкрепление агента $r(t) = R(t+1) – R(t)$ равно:

$$r(t) = \log\,[1 + u(t+1)\,\Delta X(t+1) / X(t)]. \qquad (13)$$

Для простоты предполагается, что переменная u может принимать только два значения $u = 0$ (весь капитал в деньгах) или $u = 1$ (весь капитал в акциях).

Алгоритм обучения. Система управления агента представляет собой простой адаптивный критик, состоящий из двух нейронных сетей (НС): Модель и Критик (рис. 4). Цель адаптивного критика – максимизировать функцию полезности $U(t)$:

$$U(t) = \sum_{j=0}^{\infty} \gamma^{j} r(t + j), \qquad t = 1,2,..., \qquad (14)$$

где $r(t)$ – текущее подкрепление, получаемое агентом, и γ – дисконтный фактор ($0 < \gamma < 1$).

В предположении $\Delta X(t) << X(t)$ считаем, что ситуация $\mathbf{S}(t)$, характеризующая состояние агента, зависит только от двух величин, $\Delta X(t)$ и $u(t)$: $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$.

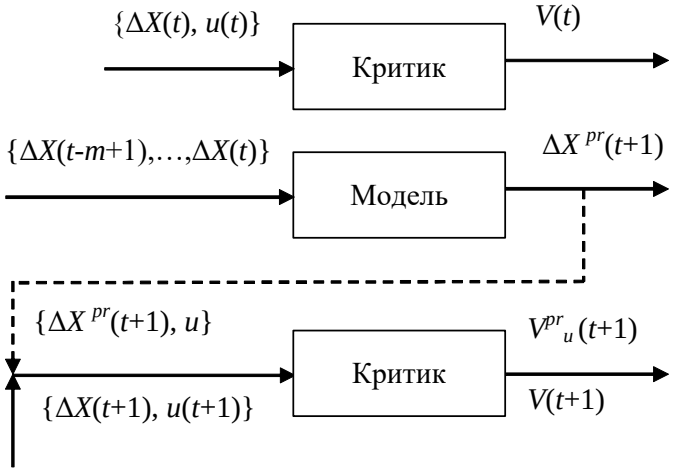

Рис. 4. Схема системы управления агента. НС Критика показана для двух последовательных тактов времени. Модель предназначена для прогнозирования изменения курса временного ряда. Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$ для текущей ситуации $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$, для ситуации в следующий такт времени $\mathbf{S}(t+1)$ и для предсказываемых ситуаций для обоих возможных действий $\mathbf{S}^{\mathbf{pr}}{}_{u}(t+1) = \{\Delta X^{pr}(t+1), u\}$, $u = 0$ либо $u = 1$.

Модель предназначена для прогнозирования изменения курса временного ряда. На вход Модели подается m предыдущих значений изменения курса $\Delta X(t-m+1)$, … , $\Delta X(t)$, на выходе формируется прогноз изменения курса в следующий такт времени $\Delta X^{pr}(t+1)$.

Модель представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x^M} = \{\Delta X(t\text{-}m\text{+}1),\ldots,\Delta X(t)\}, \quad y^M{}_j = \text{th}\,(\textstyle\sum_i w^M{}_{ij}\, x^M{}_i), \quad \Delta X^{pr}(t\text{+}1) = \textstyle\sum_j v^M{}_j\, y^M{}_j\,,$$

где $\mathbf{x^M}$ – входной вектор, $\mathbf{y^M}$ – вектор выходов нейронов скрытого слоя, $w^M{}_{ij}$ и $v^M{}_j$ – веса синапсов НС.

Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$, а именно, оценки функции полезности $U(t)$ (см. формулу (14)) для агента, находящегося в рассматриваемой ситуации $\mathbf{S}$. Критик представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x^C} = \mathbf{S}(t) = \{\Delta X(t),\, u(t)\}, \quad y^C{}_j = \text{th}\,(\textstyle\sum_{\,i}\, w^C{}_{ij}\, x^C{}_i), \quad V(t) = V(\mathbf{S}(t)) = \textstyle\sum_{\,j} v^C{}_j\, y^C{}_j\,,$$

где $\mathbf{x^C}$ – входной вектор, $\mathbf{y^C}$ – вектор выходов нейронов скрытого слоя, $w^C{}_{ij}$ и $v^C{}_j$ – веса синапсов НС.

Каждый момент времени t выполняются следующие операции:

1) Модель предсказывает следующее изменение временного ряда $\Delta X^{pr}(t\text{+}1)$.

2) Критик оценивает величину V для текущей ситуации $V(t) = V(\mathbf{S}(t))$ и для предсказываемых ситуаций для обоих возможных действий $V^{pr}{}_u(t\text{+}1) = V(\mathbf{S^{pr}}_u(t\text{+}1))$, где $\mathbf{S^{pr}}_u(t\text{+}1) = \{\Delta X^{pr}\,(t\text{+}1),\, u\}$, $u = 0$ либо $u = 1$.

3) Применяется ε-жадное правило [23]: действие, соответствующее максимальному значению $V^{pr}{}_u(t\text{+}1)$, выбирается с вероятностью $1 - \varepsilon$, и альтернативное действие выбирается с вероятностью ε ($0 < \varepsilon << 1$). Выбор действия есть выбор величины $u(t\text{+}1)$: перевести весь капитал в деньги, $u(t\text{+}1) = 0$; либо в акции, $u(t\text{+}1) = 1$.

4) Выбранное действие $u(t\text{+}1)$ выполняется. Происходит переход к моменту времени $t\text{+}1$. Подсчитывается подкрепление $r(t)$ согласно (13). Наблюдаемое значение $\Delta X(t\text{+}1)$ сравнивается с предсказанием $\Delta X^{pr}(t\text{+}1)$. Веса НС Модели подстраиваются так, чтобы минимизировать ошибку предсказания методом обратного распространения ошибки. Скорость обучения Модели равна $\alpha_M > 0$.

5) Критик подсчитывает $V(t\text{+}1) = V(\mathbf{S}(t\text{+}1))$; $\mathbf{S}(t\text{+}1) = \{\Delta X(t\text{+}1),\ u(t\text{+}1)\}$. Рассчитывается ошибка временной разности:

$$\delta(t) = r(t) + \gamma\, V(t\text{+}1) - V(t). \qquad (15)$$

6) Веса НС Критика подстраиваются так, чтобы минимизировать величину $\delta(t)$, это обучение осуществляется градиентным методом, аналогично методу обратного распространения ошибки. Скорость обучения Критика равна $\alpha_C > 0$.

Схема эволюции. Рассматривается эволюционирующая популяция, состоящая из n агентов. Каждый агент имеет ресурс $R(t)$, который изменяется в соответствии с подкреплениями агента: $R(t\text{+}1) = R(t) + r(t)$, где $r(t)$ определено в (13).

Эволюция происходит в течение ряда поколений, n_g=1,2,… Продолжительность каждого поколения n_g равна T тактов времени (T – длительность жизни агента). В начале каждого поколения ресурс каждого агента равен нулю, т.е., $R(T(n_g\text{-}1)\text{+}1) = 0$.

Начальные веса синапсов обеих НС (Модели и Критика) формируют геном агента $\mathbf{G}=\{\mathbf{W}_{M0},\mathbf{W}_{C0}\}$. Геном $\mathbf{G}$ задается в момент рождения агента и не меняется в течение его жизни. В противоположность этому текущие веса синапсов НС $\mathbf{W}_M$ и $\mathbf{W}_C$ подстраиваются в течение жизни агента путем обучения, описанного выше.

В конце каждого поколения определяется агент, имеющий максимальный ресурс $R_{max}\ (n_g)$ (лучший агент поколения n_g). Этот лучший агент порождает n потомков, которые составляют новое $(n_g\text{+}1)$-е поколение. Геномы потомков $\mathbf{G}$ отличаются от генома родителя небольшими мутациями. Более конкретно, предполагается, что в начале нового $(n_g\text{+}1)$-го поколения для каждого агента его геном формируется следующим образом $G_i(n_g\text{+}1) = G_{best,\,i}\ (n_g) + \text{rand}_i$, $\mathbf{W}_0(n_g\text{+}1) = \mathbf{G}(n_g\text{+}1)$, где $\mathbf{G_{best}}(n_g)$ – геном лучшего агента предыдущего n_g-го поколения и rand_i – нормально распределенная случайная величина с нулевым средним и стандартным отклонением

P_{mut} (интенсивность мутаций), которая добавляется к каждому весу.

Таким образом, геном **G** (начальные веса синапсов, получаемые при рождении агента) изменяется только посредством эволюции, в то время как текущие веса синапсов **W** дополнительно к этому подстраиваются посредством обучения. При этом в момент рождения агента $\mathbf{W} = \mathbf{W}_0 = \mathbf{G}$.

5.2.2. Результаты моделирования

Общие особенности адаптивного поиска. Изложенная модель была реализована в виде компьютерной программы. В компьютерных экспериментах использовалось два варианта временного ряда:

1) синусоида:

$$X(t) = 0.5(1 + \sin(2\pi t/20)) + 1, \qquad (16)$$

2) стохастический временной ряд [57]:

$$X(t) = \exp(p(t)/1200), \qquad p(t) = p(t\text{-}1) + \beta(t\text{-}1) + k_1\lambda(t), \qquad \beta(t) = k_2\beta(t\text{-}1) + \mu(t), \qquad (17)$$

где $\lambda(t)$ и $\mu(t)$ – два нормальных процесса с нулевым средним и единичной дисперсией, $k_1 = 0.3$; $k_2 = 0.9$.

Некоторые параметры модели имели одно и то же значение для всех экспериментов: дисконтный фактор $\gamma = 0.9$; количество входов НС Модели $m = 10$; количество нейронов в скрытых слоях НС Модели и Критика $N_{hM} = N_{hC} = 10$; скорость обучения Модели и Критика $\alpha_M = \alpha_C = 0.01$; параметр ε-жадного правила $\varepsilon = 0.05$; интенсивность мутаций $P_{mut} = 0.1$. Остальные параметры (продолжительность поколения T и численность популяции n) принимали разные значения в разных экспериментах, см. ниже.

Были проанализированы следующие варианты рассматриваемой модели:

– случай L (чистое обучение); в этом случае рассматривался отдельный агент, который обучался в соответствии с изложенным выше алгоритмом;
– случай E (чистая эволюция), т.е. рассматривается эволюционирующая популяция без обучения;
– случай LE (эволюция + обучение), т.е. полная модель, изложенная выше.

Было проведено сравнение ресурса, приобретаемого агентами за 200 временных тактов для этих трех способов адаптации. Для случаев E и LE бралось $T = 200$ (T – продолжительность поколения) и регистрировалось максимальное значение ресурса в популяции $R_{max}(n_g)$ в конце каждого поколения. В случае L (чистое обучение) рассматривался только один агент, ресурс которого для удобства сравнения со случаями E и LE обнулялся каждые $T = 200$ тактов времени: $R(T(n_g\text{-}1)+1) = 0$. В этом случае индекс n_g увеличивался на единицу после каждых T временных тактов, и полагалось $R_{max}(n_g) = R(T\, n_g)$.

Графики $R_{max}(n_g)$ для синусоиды (16) показаны на рис. 5. Чтобы исключить уменьшение значения $R_{max}(n_g)$ из-за случайного выбора действий при применении ε-жадного правила для случаев LE и L, полагалось $\varepsilon = 0$ после $n_g = 100$ для случая LE и после $n_g = 2000$ для случая L (на рис. 5 видно резкое увеличение $R_{max}(n_g)$ после $n_g = 100$ и $n_g = 2000$ для соответствующих случаев). Результаты усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

Рис. 5 показывает, что обучение, объединенное с эволюцией (случай LE), и чистая эволюция (случай E) дают одно и то же значение конечного ресурса $R_{max}(500) = 6.5$. Однако эволюция и обучение вместе обеспечивают нахождение больших значений R_{max} быстрее, чем эволюция отдельно – существует симбиотическое взаимодействие между обучением и эволюцией.

Из (13) следует, что существует оптимальная стратегия поведения агента: вкладывать весь капитал в акции ($u(t+1) = 1$) при прогнозе роста курса ($\Delta X(t+1) > 0$), вкладывать весь капитал в деньги ($u(t+1) = 0$) при прогнозе падения курса ($\Delta X(t+1) < 0$). Анализ экспериментов, представленных на рис. 5, показал, что в случаях LE (обучение + эволюция), и E (чистая эволюция) такая оптимальная стратегия находится. Это соответствует асимптотическому значению ресурса $R_{max}(500) = 6.5$.

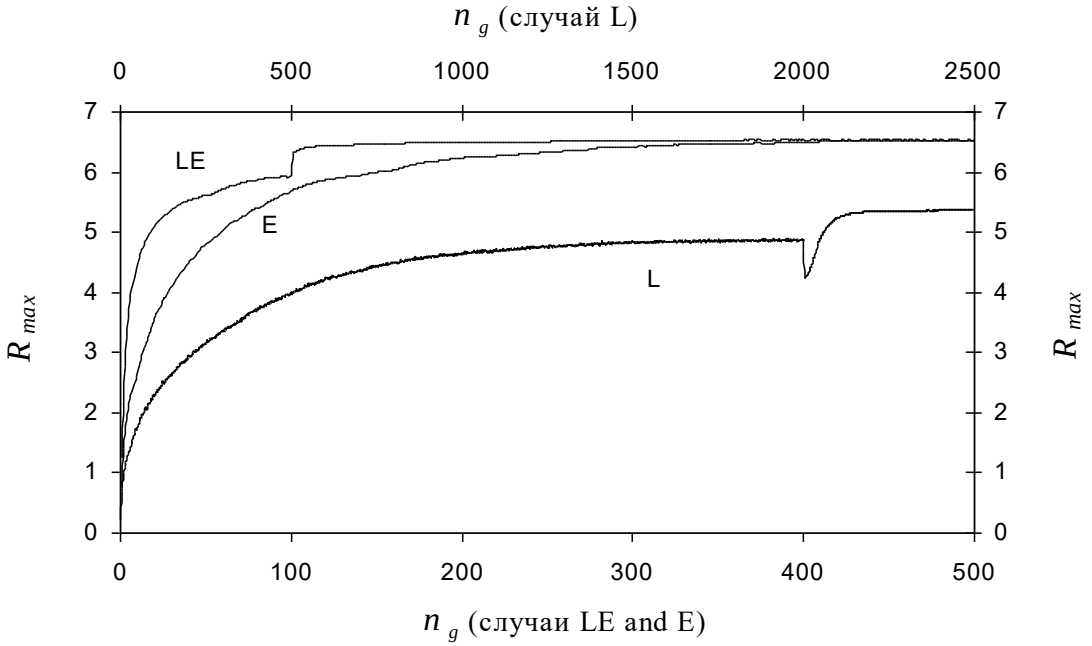

Рис. 5. Зависимости $R_{max}(n_g)$. Кривая LE соответствует случаю эволюции, объединенной с обучением, кривая E – случаю чистой эволюции, кривая L – случаю чистого обучения. Временная шкала для случаев LE и E (номер поколения n_g) представлена снизу, для случая L (индекс n_g) – сверху. Моделирование проведено для синусоиды, кривые усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

В случае L (чистое обучение) асимптотическое значение ресурса ($R_{max}(2500) = 5.4$) существенно меньше. Анализ экспериментов для этого случая показал, что одно обучение обеспечивает нахождение только следующей «субоптимальной» стратегии поведения: агент держит капитал в акциях при росте и при слабом падении курса и переводит капитал в деньги при сильном падении курса. Та же тенденция к явному предпочтению вкладывать капитал в акции при чистом обучении наблюдалась и для экспериментов на стохастическом ряде (17).

Взаимодействие между обучением и эволюцией. Эффект Болдуина. Рис. 5 демонстрирует, что поиск оптимальной стратегии посредством только эволюции происходит медленнее, чем при эволюции, объединенной с обучением (см. кривые E и LE на этом рисунке). Хотя обучение в данной модели само по себе не оптимально, оно помогает эволюции находить лучшие стратегии.

Если длительность поколения T была достаточно большой (1000 и более тактов времени), то для случая LE часто наблюдалось и более явное влияние обучения на эволюционный процесс. В первых поколениях эволюционного процесса существенный рост ресурса агентов наблюдался не с самого начала поколения, а спустя 200-300 тактов, т.е. агенты явно обучались в течение своей жизни находить более или менее приемлемую стратегию поведения, и только после смены ряда поколений рост ресурса начинался с самого начала поколения. Это можно интерпретировать как проявление известного эффекта Болдуина: исходно приобретаемый навык в течение ряда

поколений становился наследуемым [59, 60]. Этот эффект наблюдался в ряде экспериментов, один из которых представлен на рис. 6. Для этого эксперимента было проанализировано, как изменяется значение ресурса наилучшего агента в популяции $R_{max}(t)$ в течение первых пяти поколений. Расчет был проведен для синусоидального ряда (16). Рис. 6 показывает, что в течение первых двух поколений значительный рост ресурса лучшего в популяции агента начинается только после задержки 100-300 тактов времени; т.е., очевидно, что агент оптимизирует свою стратегию поведения при помощи обучения. От поколения к поколению агент находит хорошую стратегию поведения все раньше и раньше. К пятому поколению лучший агент «знает» хорошую стратегию поведения с самого рождения, и обучение не приводит к существенному улучшению стратегии. Итак, рис. 6 показывает, что стратегия, изначально приобретаемая посредством обучения, становится наследуемой (эффект Болдуина).

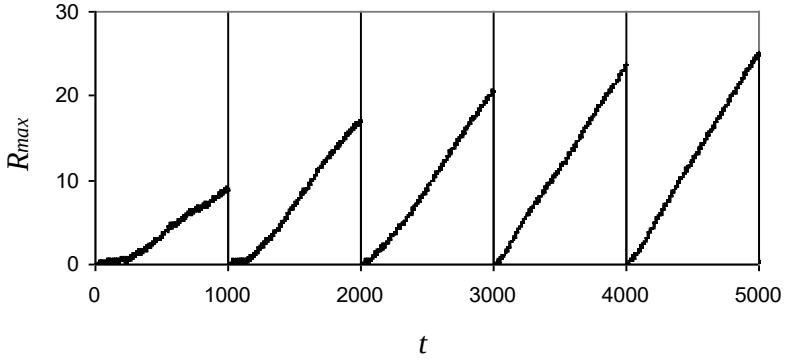

Рис. 6. Зависимость ресурса лучшего в популяции агента R_{max} от времени t для первых пяти поколений. Случай LE (эволюция, объединенная с обучением); размер популяции n = 10, длительность поколения T = 1000. Моменты смены поколений показаны вертикальными линиями. Для первых двух поколений есть явная задержка в 100-300 тактов времени в росте ресурса агента. К пятому поколению лучший агент «знает» хорошую стратегию поведения с самого рождения, т.е. стратегия, изначально приобретаемая посредством обучения, становится наследуемой.

Были проанализированы различные наборы параметров модели и выяснено, что эффект Болдуина стабильно проявляется, если продолжительность поколения T составляет 1000 и более тактов времени, что обеспечивает достаточно эффективное обучение в течение жизни агента.

Особенности предсказания Модели. Практика не есть критерий истины. Система управления агента включает в себя нейронную сеть Модели, предназначенную для предсказания изменения значения $\Delta X(t+1)$ временного ряда в следующий такт времени $t+1$. Анализ работы Модели обнаружил очень интересную особенность. Нейронная сеть Модели может давать неверные предсказания, однако агент, тем не менее, может использовать эти предсказания для принятия верных решений – практика не есть критерий истины. Например, рис. 7 показывает предсказываемые изменения $\Delta X^{pr}(t+1)$ и реальные изменения $\Delta X(t+1)$ стохастического временного ряда в случае чистой эволюции (случай E). Предсказания нейронной сети Модели достаточно хорошо совпадают по форме с кривой $\Delta X(t)$. Однако, предсказанные значения $\Delta X^{pr}(t+1)$ отличаются примерно в 25 раз от значений $\Delta X(t+1)$.

На рис. 8 приведен другой пример особенностей предсказания нейронной сети Модели в случае LE (эволюция, объединенная с обучением). Этот пример показывает, что предсказания нейронной сети Модели могут отличаться от реальных данных не только масштабом, но и знаком.

Хотя предсказания Модели могут быть неверными количественно, можно предположить, что правильность их формы или правильность после линейных преобразований (например, изменения знака) приводит к тому, что Модель является

полезной для адаптивного поведения. Эти предсказания эффективно используются системой управления агентов для нахождения оптимальной поведения: стратегия поведения агентов для обоих приведенных примеров работы Модели была практически оптимальна.

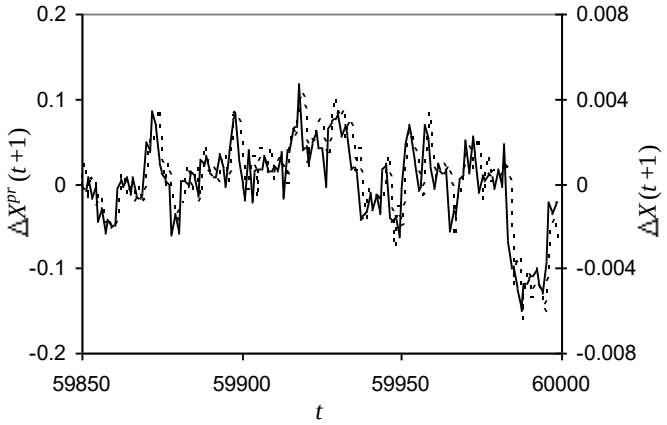

Рис. 7. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай чистой эволюции. $n = 10$, $T = 200$. Хотя обе кривые имеют сходную форму, по величине $\Delta X^{pr}(t+1)$ и $\Delta X(t+1)$ радикально различаются.

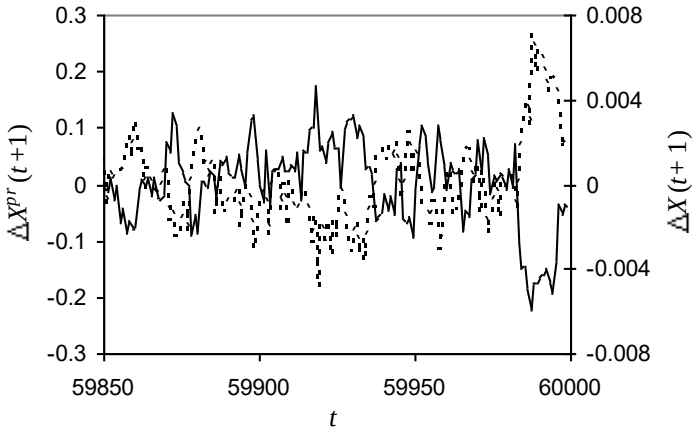

Рис. 8. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай эволюции, объединенной с обучением. $n = 10$, $T = 200$. Кривые $\Delta X^{pr}(t+1)$ и $\Delta X(t+1)$ различаются как по величине, так и знаком.

По-видимому, наблюдаемое увеличение значений ΔX^{pr} нейронной сетью Модели полезно для работы нейронной сети Критика, так как реальные значения $\Delta X(t+1)$ слишком малы (порядка 0.001). Таким образом, нейронная сеть Модели может не только предсказывать значения $\Delta X^{pr}(t+1)$, но также осуществлять полезные преобразования этих значений.

Эти особенности работы нейронной сети Модели обусловлены доминирующей ролью эволюции над обучением при оптимизации системы управления агентов. На самом деле, из-за малой длительности поколений ($T = 200$) в проведенном моделировании, веса синапсов нейронных сетей изменяются большей частью за счет эволюционных мутаций. Такой процесс делает предпочтительными такие системы управления, которые устойчивы в эволюционном смысле. Кроме того, важно подчеркнуть, что задача, которую «решает» эволюция в рассматриваемой модели, значительно проще, чем та задача, которую «решает» обучение. Эволюции достаточно обеспечить выбор из двух действий, приводящий к награде. А схема обучения

предусматривает довольно сложную процедуру прогноза ситуации **S**, оценки качества прогнозируемых ситуаций, итеративного формирования оценок качества ситуаций $V(\mathbf{S})$ и выбора действия на основе этих оценок. То есть эволюция идет к нужному результату более прямым путем, а так как задача агентов проста, то эволюция в определенной степени «задавливает» довольно сложный механизм обучения. Тем не менее, есть определенная синергия во взаимодействии обучения и эволюции: обучение ускоряет процесс поиска оптимальной стратегии поведения.

Опыт работы с исследованной моделью показывает, что необходима определенная осторожность в выборе базовой модели функциональной системы для проекта «Мозг анимата». Имеет смысл рассмотреть и другие возможности для функциональной системы. Например, в [61] начата проработка новой версии «Мозга анимата» на основе использования понятия хеббовских ансамблей [62]. Однако проведенный в [61] анализ показывает, что необходимость согласования всех элементов системы управления аниматом накладывает довольно серьезные ограничения на ее архитектуру, и важны дальнейшие исследования по данному перспективному проекту, в том числе и на основе опыта разработки представленной выше архитектуры на базе нейросетевых адаптивных критиков.

6. КОНТУРЫ ПРОГРАММЫ БУДУЩИХ ИССЛЕДОВАНИЙ

Анализ современных исследований адаптивного поведения [24,47] показывает, что хотя проделана большая работа и есть много интересных отдельных моделей, ученые еще далеки от понимания того, как возникали и развивались системы управления живых организмов, как развитие этих систем способствовало эволюции когнитивных способностей животных, и как процесс когнитивной эволюции привел к возникновению интеллекта человека. Образно говоря, у нас есть некоторые небольшие фрагменты картины, но мы еще видим всей картины. Есть определенные подходы к исследованиям, но само исследование интеллектуального адаптивного поведения, природы естественного интеллекта, эволюционного происхождения интеллекта еще не проведено. Предложим эскизный план будущих исследований, направленных на анализ проблемы происхождения интеллекта.

А) Разработка схем и моделей адаптивного поведения анимата на базе проекта «Мозг анимата». В разделе 5 изложен проект «Мозг анимата», который предложен как общая «платформа» для систематизированного построения широкого спектра моделей адаптивного поведения. Реализация в моделях схем и конструкций «Мозга анимата» для анимата, обладающего естественными потребностями (питания, размножения, безопасности) могла бы стать первым и важным шагом планируемых исследований.

Б) Исследование перехода от физического уровня обработки информации в нервной системе животных к уровню обобщенных образов. Такой переход можно рассматривать, как появление в «сознании» животного свойства «понятие». Обобщенные образы можно представить как мысленные аналоги наших слов, не произносимых животными, но реально используемых ими. Например, у собаки явно есть понятия «хозяин», «свой», «чужой», «пища». И важно осмыслить, как такой весьма нетривиальный переход мог произойти в процессе эволюции.

В) Исследование процессов формирования причинных связей в памяти животных. По-видимому, запоминание причинно-следственных связей между событиями во внешней среде и адекватное использование этих связей в поведении – одно из ключевых свойств активного познания животным закономерностей внешнего мира. Такая связь формируется, например, при выработке условного рефлекса: животное запоминает связь между условным стимулом (УС) и следующим за ним безусловным стимулом (БС), что позволяет ему предвидеть события в окружающем мире и адекватно использовать это предвидение. Естественный следующий шаг –

переход от отдельных причинных связей к логическим выводам на основе уже сформировавшихся знаний.

Г) Исследование процессов формирования логических выводов в «сознании» животных. Фактически, уже на базе классического условного рефлекса животные способны делать «логический вывод» вида: {УС, УС --> БС} => БС или «Если имеет место условный стимул, и за условным стимулом следует безусловный, то нужно ожидать появление безусловного стимула». Можно говорить, что такие выводы подобны выводам математика, доказывающего теоремы (раздел 2). И целесообразно разобраться в системах подобных выводов, понять, насколько адаптивна логика поведения животных и насколько она подобна нашей, человеческой логике.

Д) Исследование коммуникаций, возникновения языка. Наше мышление тесно связано с языком, с языковым общением между людьми. Поэтому целесообразно проанализировать: как в процессе биологической эволюции возникал язык общения животных, как развитие коммуникаций привело к современному языку человека, как развитие коммуникаций и языка способствовало развитию логики, мышления, интеллекта человека.

Конечно же, перечисленные пункты формируют только контуры плана будущих исследований. Тем не менее, уже сейчас видно, сколь широк фронт исследований, и как много нетривиальной, интересной и важной работы предстоит сделать.

Актуальность исследования проблемы происхождения интеллекта подчеркивает необходимость формирования серьезной академической научной программы «Эволюция, Мозг, Интеллект», в рамках которой велись бы работы по моделям адаптивного поведения.

СПИСОК ЛИТЕРАТУРЫ

1. Эйген М. *Самоорганизация материи и эволюция биологических макромолекул*. М.: Мир. 1973. 216 с.
2. Эйген М., Шустер П. *Гиперцикл. Принципы самоорганизации макромолекул*. М.: Мир. 1982. 270 с.
3. Eigen M., Gardiner W., Schuster P., Winkler-Oswatich R. The origin of genetic information. *Scientific American.* 1981. **244(**4). 88-118.
4. Ратнер В.А., Шамин В.В. Сайзеры: моделирование фундаментальных особенностей молекулярно-биологической организации. Соответствие общих свойств и конструктивных особенностей коллективов макромолекул. *Журн. общ. биологии*. 1983. **44**(1). 51-61.
5. Редько В.Г. *Эволюционная кибернетика*. М.: Наука. 2001. 156 с.
6. Редько В.Г. *Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики.* Серия «Синергетика: от прошлого к будущему». М.: УРСС. 2005. 224 с.
7. Воронин Л.Г. *Эволюция высшей нервной деятельности*. М.: Наука. 1977. 128 с.
8. *From animals to animats.* Proceedings of the First International Conference on Simulation of Adaptive Behavior**.** Eds. Meyer J.-A., Wilson S. W. Cambridge, Massachusetts, London, England: The MIT Press. 1990.
9. Meyer J.-A., Guillot, A. From SAB90 to SAB94: *Four years of Animat research.* In: Proceedings of the Third International Conference on Simulation of Adaptive Behavior**.** Eds. Cliff, Husbands, Meyer J.-A., Wilson S.W. Cambridge: The MIT Press: 1994. http://animatlab.lip6.fr/index.en.html
10. Guillot A., Meyer J.-A. From SAB94 to SAB2000: What's new, Animat? In: *From Animals to Animats 6.* Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior. Eds. Meyer et al. The MIT Press. 2000. http://animatlab.lip6.fr/index.en.html

11. Непомнящих В.А. Аниматы как модель поведения животных. *IV Всероссийская научно-техническая конференция «Нейроинформатика-2002».* Материалы дискуссии «Проблемы интеллектуального управления – общесистемные, эволюционные и нейросетевые аспекты». М.: МИФИ. 2003. с. 58-76.
12. Непомнящих В.А. Поиск общих принципов адаптивного поведения живых организмов и аниматов. *Новости искусственного интеллекта*. 2002. **2**. 48-53.
13. Donnart J.Y., Meyer J.A. Learning reactive and planning rules in a motivationally autonomous animat. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics.* 1996. **26**(3). pp.381-395. http://animatlab.lip6.fr/index.en.html
14. Wilson S.W. *The animat path to AI*. In: *From animals to animats.* Proceedings of the First International Conference on Simulation of Adaptive Behavior**.** Eds. Meyer J.-A., Wilson S. W. Cambridge, Massachusetts, London, England: The MIT Press. 1990. pp. 15-21.
15. Цетлин М.Л. *Исследования по теории автоматов и моделирование биологических систем*. М.: Наука. 1969. 316 с.
16. Варшавский В.И., Поспелов Д.А. *Оркестр играет без дирижера*. М.: Наука. 1984.
17. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – «Животное». *Моделирование обучения и поведения.* М.: Наука. 1975. с.152-171.
18. Гаазе-Рапопорт М.Г., Поспелов Д.А. *От амебы до робота: модели поведения*. М.: Наука. 1987.
19. Holland J.H. *Adaptation in Natural and Artificial Systems*. Boston, MA: MIT Press. 1992.
20. Курейчик В.М. *Генетические алгоритмы и их применение*. Таганрог: ТРТУ. 2002.
21. Емельянов В.В., Курейчик В.М., Курейчик В.В. *Теория и практика эволюционного моделирования*. М.: Физматлит. 2003.
22. Holland J.H., Holyoak K.J., Nisbett R.E., Thagard P. *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press. 1986.
23. Sutton R., Barto A. *Reinforcement Learning: An Introduction.* Cambridge: MIT Press. 1998. http://www.cs.ualberta.ca/~sutton/book/the-book.html
24. *From animals to animats 9*. Proceedings of the Ninth International Conference on Simulation of Adaptive Behaviour. Eds. Nolfi S., Baldassarre G., Calabretta R., Hallam J., Marocco D., Miglino O., Meyer J-A, Parisi D. Berlin, Germany: Springer Verlag. 2006. **4095**.
25. Сайт AnimatLab: http://animatlab.lip6.fr/index.en.html
26. Сайт AI Laboratory of Zurich University: http://www.ifi.unizh.ch/groups/ailab/
27. Pfeifer R., Scheier C. *Understanding Intelligence*. MIT Press. 1999.
28. Сайт Laboratory of Artificial Life and Robotics: http://gral.ip.rm.cnr.it/
29. Nolfi S., Floreano D. *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines.* Cambridge, MA: MIT Press/Bradford Books. 2000. 384 p.
30. Сайт MIT Computer Science and Artificial Intelligence Laboratory: http://www.csail.mit.edu/index.php
31. Brooks R.A. *Cambrian Intelligence: The Early History of the New AI*. MIT Press. 1999.
32. Сайт Neuroscience Institute: http://www.nsi.edu/
33. Krichmar J.L., Edelman G.M. Machine psychology: autonomous behavior, perceptual categorization and conditioning in a brain-based device. *Cerebral Cortex*. 2002. **12**. 818-830.
34. Krichmar J.L., Edelman G.M. Brain-based devices: intelligent systems based on principles of the nervous system. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV. 2003. pp. 940-945.
35. Krichmar J.L., Seth A.K., Nitz D.A., Fleischer J.G., Edelman G.M. Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions. *Neuroinformatics*. 2005. **3**(3). 197-221. http://www.nsi.edu/nomad/pubs.html http://www.nsi.edu/nomad/krichmar_neuroinf_2005.pdf

36. Непомнящих В.А. Как животные решают плохо формализуемые задачи поиска. Синергетика и психология: Тексты. Выпуск 3: *Когнитивные процессы*. Ред. Аршинов В.И., Трофимова И.Н., Шендяпин В.М. М.: Когито-Центр. 2004. С.197-209.
37. Nepomnyashchikh V.A., Podgornyj K.A. Emergence of adaptive searching rules from the dynamics of a simple nonlinear system. *Adaptive Behavior*. 2003. **11**(4). 245-265.
38. Жданов А.А. Метод автономного адаптивного управления. *Изв. РАН. Серия Теория и системы управления.* 1999. **5**. 127-134.
39. Жданов А.А. О методе автономного адаптивного управления. В сб.: *VI Всероссийская научно-техническая конференция «Нейроинформатика-2004». Лекции по нейроинформатике*. Часть 2. М.: МИФИ. 2004. с. 15-56.
40. Станкевич Л.А. Нейрологические средства систем управления интеллектуальных роботов. В сб.: *VI Всероссийская научно-техническая конференция «Нейроинформатика-2004». Лекции по нейроинформатике*. Часть 2. М.: МИФИ. 2004. С. 57-110.
41. Самарин А.И. Модель адаптивного поведения мобильного робота, реализованная с использованием идей самоорганизации нейронных структур. В сб.: *IV Всероссийская научно-техническая конференция «Нейроинформатика-2002». Материалы дискуссии «Проблемы интеллектуального управления – общесистемные, эволюционные и нейросетевые аспекты»*. М.: МИФИ. 2003. с. 106-120.
42. Бурцев М.С., Гусарев Р.В., Редько В.Г. Исследование механизмов целенаправленного адаптивного управления. *Изв. РАН. Серия Теория и системы управления.* 2002. **6**. 55-62.
43. Бурцев М.С. *Модель эволюционного возникновения целенаправленного адаптивного поведения. 2. Исследование развития иерархии целей*. Препринт ИПМ РАН. 2002. **69**.
44. Burtsev M.S., Turchin P.V. Evolution of cooperative strategies from first principles. *Nature*. 2006. **440**(7087). 1041-1044.
45. Мосалов О.П., Редько В.Г. Непомнящих В.А. *Модель поискового поведения анимата*. Препринт ИПМ РАН. 2003. **19**.
46. Red'ko V.G., Mosalov O.P., Prokhorov D.V. A Model of Evolution and Learning. *Neural Networks*. 2005. **18**(5-6). 738-745.
47. *От моделей поведения к искусственному интеллекту.* Серия «Науки об искусственном» (под ред. Редько В.Г.). М.: УРСС. 2006.
48. *Learning and Approximate Dynamic Programming: Scaling Up to the Real World*. Eds. Jennie Si, Andrew Barto, Warren Powell, and Donald Wunsch. IEEE Press and John Wiley & Sons. 2004.
49. Rumelhart D.E., Hinton G.E., Williams R.G. Learning representation by back-propagating error. *Nature*. 1986. **323**(6088). 533-536.
50. Редько В.Г., Прохоров Д.В. Нейросетевые адаптивные критики. В сб.: *VI Всероссийская научно-техническая конференция «Нейроинформатика-2004». Сборник научных трудов*. Часть 2. М.: МИФИ. 2004. с.77-84.
51. Prokhorov D.V., Wunsch D.C. Adaptive critic designs. *IEEE Trans. Neural Networks*. 1997. **8**(5). 997-1007.
52. Prokhorov D.V. Backpropagation through time and derivative adaptive critics: a common framework for comparison. In: *Learning and Approximate Dynamic Programming: Scaling Up to the Real World*. Eds. Jennie Si, Andrew Barto, Warren Powell, and Donald Wunsch. IEEE Press and John Wiley & Sons. 2004. http://mywebpages.comcast.net/dvp/
53. Анохин К.В., Бурцев М.С., Зарайская И.Ю., Лукашев А.О., Редько В.Г. Проект «Мозг анимата»: разработка модели адаптивного поведения на основе теории функциональных систем. В сб.: *Восьмая национальная конференция по*

искусственному интеллекту с международным участием. Труды конференции. М.: Физматлит. 2002. **2**. с.781-789.
54. Red'ko V.G., Prokhorov D.V., Burtsev M.S. Theory of functional systems, adaptive critics and neural networks. In: *International Joint Conference on Neural Networks*, Budapest. 2004. pp. 1787-1792.
55. Анохин П.К. *Системные механизмы высшей нервной деятельности*. М.: Наука. 1979. 453 с.
56. Red'ko V.G., Mosalov O.P., Prokhorov D.V. A model of Baldwin effect in populations of self-learning agents. In: *International Joint Conference on Neural Networks*, Montreal. 2005.
57. Prokhorov D., Puskorius G., Feldkamp L. Dynamical neural networks for control. In: *A field guide to dynamical recurrent networks.* Eds. J. Kolen and S. Kremer. NY: IEEE Press. 2001. pp. 257-289.
58. Moody J., Wu L., Liao Y., Saffel M. Performance function and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*. 1998. **17**. 441-470.
59. Baldwin J.M. A new factor in evolution. *American Naturalist*. 1896. **30**. 441-451.
60. *Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect. Special Issue of Evolutionary Computation on the Baldwin Effect.* Eds. Turney P., Whitley D., Anderson R. 1996. **4**(3).
61. Red'ko V.G., Anokhin K.V., Burtsev M.S., Manolov A.I., Mosalov O.P., Nepomnyashchikh V.A., Prokhorov D.V. Project “Animat Brain”: Designing the Animat Control System on the Basis of the Functional Systems Theory. In: *The Ninth International Conference on the Simulation of Adaptive Behavior (SAB'06)*, 25 - 29 September 2006, CNR, Roma, Italy, Third Workshop on Anticipatory Behavior in Adaptive Learning Systems (ABiALS 2006), Proceedings.
62. Hebb D.O. *The Organization of Behavior. A Neuropsychological Theory*. N.Y.: Wiley & Sons. 1949. 355 p.

Материал поступил в редакцию 24.05.2007, опубликован 07.06.2007.