ПРОБЛЕМА ПРОИСХОЖДЕНИЯ ИНТЕЛЛЕКТА И МОДЕЛИ АДАПТИВНОГО ПОВЕДЕНИЯ*

В.Г. Редько
Институт оптико-нейронных технологий РАН
E-mail: redko@iont.ru

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 адаптивных агентов на основе адаптивных критиков. В порядке обсуждения предлагается программа будущих исследований эволюции адаптивного поведения.

PROBLEM OF EVOLUTIONARY ORIGIN OF INTELLIGENCE AND MODELS OF ADAPTIVE BEHAVIOR

V.G. Red'ko
Institute of Optical Neural Technologies, RAS
E-mail: redko@iont.ru

Abstract. The approach to the problem of evolutionary origin of intelligence is discussed; the approach is based on modeling of evolution of adaptive behavior. The works of leading laboratories in the field of simulation of adaptive behavior are characterized. A special attention is paid to reinforcement learning and adaptive critic designs. The project “Animat Brain” directed to development of a general platform for systematic designing of models of adaptive behavior is described. The results of the concrete model of evolution of self-learning agents that are based of adaptive critic designs are represented. The sketch program for future research of evolution of adaptive behavior is proposed.

1. Можно ли обосновать математику?

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.

И, 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

* Работа выполнена финансовой поддержке программы Президиума РАН "Интеллектуальные компьютерные системы" (проект 2-45) и РФФИ (проект № 04-01-00179).

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. По крайней мере, можно попытаться это сделать. Почему можно ожидать положительный ответ на этот вопрос? А давайте попробуем 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.

Мы можем пойти и дальше – можем представить, что в памяти собаки есть *семантическая сеть*, сеть связей между понятиями, образами. Например, мы можем представить, что у собаки есть понятия "пища", "опасность", "другая собака". С понятием "пища" могут быть связаны понятия "мясо", "косточка". При выработке пищевого условного рефлекса, например, на звонок (скажем, УС = "звонок", БС = "мясо") у собаки, по-видимому, формируется простая семантическая связь (рис.1).

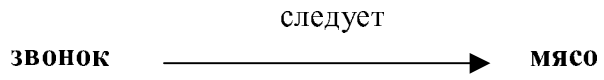

Рис.1. Гипотетическая семантическая связь, формируемая в памяти собаки.

Далее можно представить процессы формирования разнообразных семантических сетей в процессе жизни собаки и накопления ей жизненного опыта. Такие семантические сети, формируемые в памяти собаки, по-видимому, аналогичны семантическим сетям, исследуемым разработчиками искусственного интеллекта [1].

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 (рис. 2) [2]. То есть, целесообразно с помощью моделей представить общую картину эволюции когнитивных способностей животных и эволюционного происхождения интеллекта человека.

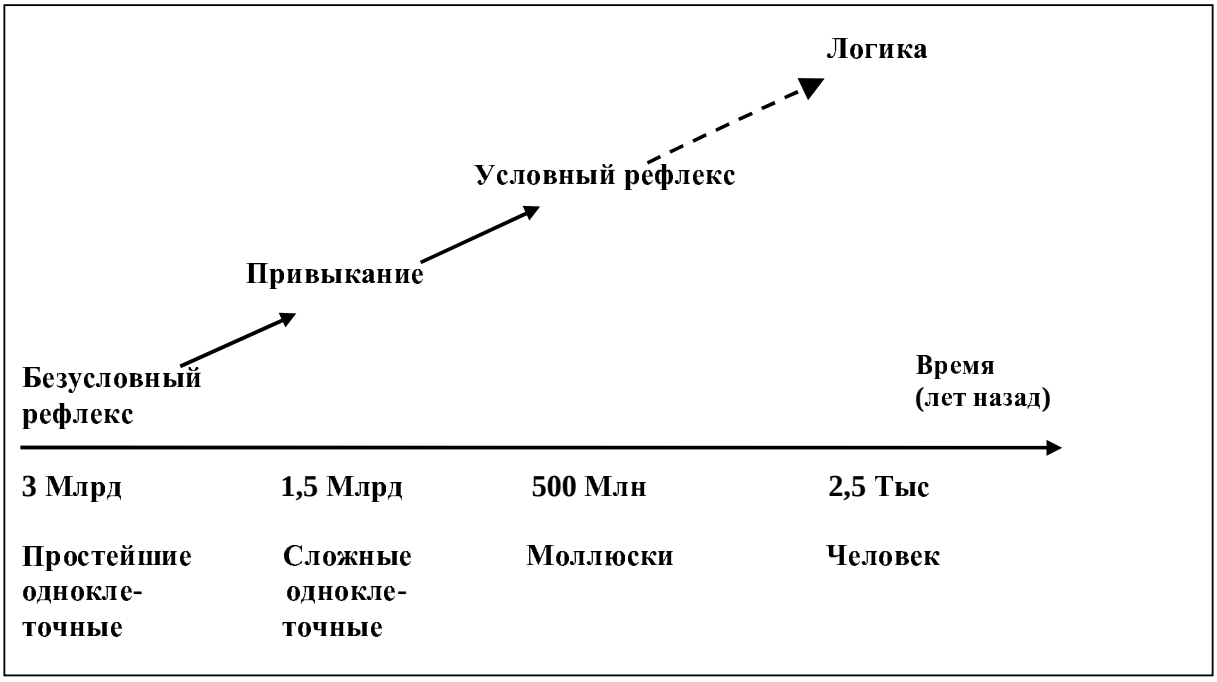

Рис. 2. "Интеллектуальные изобретения" биологической эволюции. "Авторы изобретений" и "даты приоритетов" представлены довольно условно.

Естественно, что такие исследования – это огромный фронт работы, и задачу построения теории происхождения мышления можно пока рассматривать как сверхзадачу. Разработка такой теории могла бы обеспечить определенное обоснование применимости нашего мышления в научном познании, то есть, укрепить фундамент всего величественного здания науки. Чтобы вести эту работу серьезно, целесообразно идти именно по пути построения математических и компьютерных моделей когнитивной эволюции.

Есть ли задел таких исследований? Оказывается, что да, есть. Сравнительно недавно сформировалось направление исследований «Адаптивное поведение», дальняя цель которого очень близка к задаче построения теории происхождения мышления. В следующем разделе мы обсудим модели адаптивного поведения.

2. Модели адаптивного поведения

2.1. From Animal to Animat – модели адаптивного поведения животного и робота

Направление "Адаптивное поведение" (АП) активно развивается с начала 1990-х годов [3-5]. 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). Также часто используется термин "агент".

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* [6,7].

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 [8].

Программа-максимум близка к очерченной выше задаче построения теории происхождения мышления.

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 [9].

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" [10,11].

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 [12,13].

Хороший обзор ранних работ по адаптивному поведению, представлен в книге М.Г. Гаазе-Рапопорта, Д.А. Поспелова "От амебы до робота: модели поведения" [13].

В исследованиях АП используется ряд нетривиальных компьютерных методов:
- нейронные сети,
- генетический алгоритм и другие методы эволюционной оптимизации [14-17],
- классифицирующие системы (Classifier Systems) [18],
- обучение с подкреплением (Reinforcement Learning) [19].

Метод обучения с подкреплением будет кратко охарактеризован в разделе 2.3.

Подчеркнем, что в АП в основном используется *феноменологический подход* к исследованиям систем управления адаптивным поведением. Т.е. 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 должен быть более эффективен (по крайней мере, на начальных этапах работ), так как очень трудно сформировать целостную картину поведения на

основе анализа всего сложного многообразия функционирования нейронов, синапсов, молекул.

2.2. Исследователи адаптивного поведения

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

- AnimatLab (Париж, руководитель – один из инициаторов данного направления Жан-Аркадий Мейер) [3,8,20]. В этой лаборатории ведется широкий спектр исследований адаптивных роботов и адаптивного поведения животных. Подход AnimatLab предполагает, что система управления анимата может формироваться и модифицироваться посредством 1) *обучения*, 2) индивидуального *развития* (онтогенеза) и 3) *эволюции*.
- Лаборатория искусственного интеллекта в университете Цюриха (руководитель Рольф Пфейфер) [21,22]. Основной подход этой лаборатории – познание природы интеллекта путем его создания ("understanding by building"). Он включает в себя 1) построение моделей биологических систем, 2) исследование общих принципов естественного интеллекта животных и человека, 3) использование этих принципов при конструировании роботов и других искусственных интеллектуальных систем.
- Лаборатория искусственной жизни и роботики в Институте когнитивных наук и технологий (Рим, руководитель Стефано Нолфи) [23,24], ведущая исследования в области эволюционной роботики и принципов формирования адаптивного поведения.
- Лаборатория информатики и искусственного интеллекта в Массачусетском технологическом институте (руководитель Родни Брукс) [25,26], 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 [27-29].

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

- модели поискового адаптивного поведения на основе спонтанной активности, приводящей к переключениям между разными тактиками поиска, например, тактикой движения по градиенту источника запаха (при поиске пищи) и тактикой случайного поиска [6,7,30,31] (В.А. Непомнящих, Институт биологии внутренних вод им. И.Д. Папанина РАН);
- концепции и модели автономного адаптивного управления на основе аппарата эмоций [32,33] (А.А. Жданов, Институт системного программирования РАН);
- разработку принципов построения систем управления антропоморфных и гуманоидных роботов [34] (Л.А. Станкевич, Санкт-Петербургский политехнический университет);
- разработку нейросетевых моделей поведения роботов и робототехнических устройств [35] (А.А. Самарин, НИИ нейрокибернетики им. А.Б. Когана РГУ);
- модели адаптивного поведения на основе эволюционных и нейросетевых методов, в частности, модели эволюционного возникновения целенаправленного адаптивного поведения [36-40] (В.Г. Редько, М.С. Бурцев, О.П. Мосалов, Институт оптико-нейронных технологий РАН, Институт прикладной математики им. М.В. Келдыша РАН).

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.

2.3. Обучение 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 [19].

Идейным вдохновителем этих работ был А.Г. Клопф (Air Force, USA), который в книге "Целеустремленный нейрон" предложил несколько спорную, но достаточно четкую и последовательную методологию исследований памяти, обучения, адаптивного поведения [41].

Общая схема обучения с подкреплением [19] показана на рис. 3.

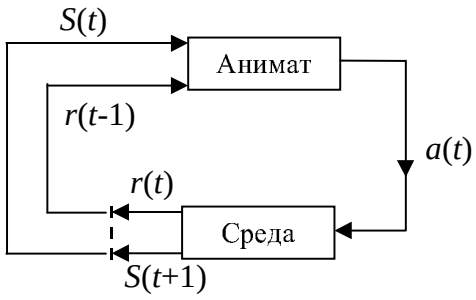

Рис. 3. Схема обучения с подкреплением.

Рассматривается анимат, взаимодействующий с внешней средой. Время предполагается дискретным: $t = 1,2,\ldots$ В текущей ситуации анимат $S(t)$ выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $S(t+1)$. Подкрепление может быть положительным (награда) или отрицательным (наказание).

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 коэффициента забывания:

$$U(t) = \sum_{k=0}^{\infty} \gamma^k \, r(t+k) \quad , \qquad (1)$$

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

В процессе обучения анимат формирует *политику* (стратегию поведения). Политика определяет выбор (детерминированный или вероятностный) действия в зависимости от ситуации. Р. Саттон и Э. Барто [19] исследовали ряд методов формирования политики, основанных на динамическом программировании и методах Монте-Карло.

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

2.3.1. 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:

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

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 [19]:

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

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, $0 < \varepsilon << 1$.

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.

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 [42].

Важное достоинство метода обучения с подкреплением – его простота. То есть 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, 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 [19]) и был использован многочисленных приложениях. В частности, применения этого метода включают в себя:
- оптимизацию игры в триктрак (достигнут уровень мирового чемпиона);
- оптимизацию системы управления работы лифтов;
- формирование динамического распределения каналов для мобильных телефонов;
- оптимизацию расписания работ на производстве.

Подчеркнем, что метод обучения с подкреплением может рассматриваться как развитие автоматной теории адаптивного поведения, разработанной в работах М.Л. Цетлина и его последователей [10,11].

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.

2.3.2. Нейросетевые адаптивные критики [43]

Конструкции адаптивных критиков можно рассматривать как развитие моделей обучения с подкреплением на случай, когда как ситуации, так и действия задаются векторами **S** и **A** и изложенная выше схема итеративного формирования матрицы $Q(S_j, a_i)$ не работает. В этом случае 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 случае аппроксимации с помощью нейронных сетей, параметрами аппроксимирующих функций являются веса синапсов нейросети, оптимизация производится путем подстройки весов, например, аналогично тому, как это делается в методе обратного распространения ошибки.

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

Критик – это блок системы управления, который оценивает 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. 4. Предполагаем, что как Критик, так и Контроллер представляют собой многослойные персептроны (такие же, какие используются в методе обратного распространения ошибки) с весами синапсов $\mathbf{W_C}$ и $\mathbf{W_A}$, соответственно.

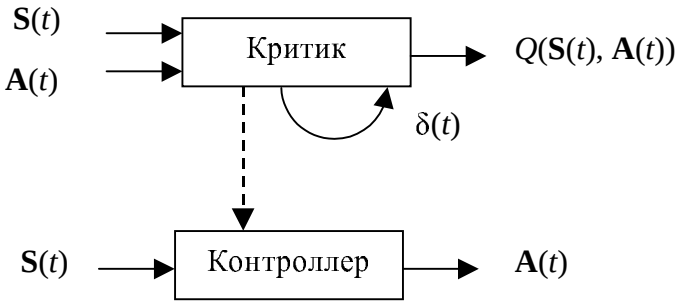

Рис. 4. Схема Q-критика.

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 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 [19]. 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. 5.

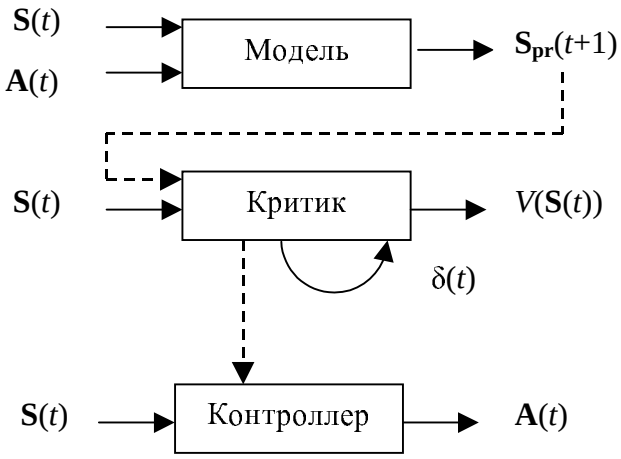

Рис. 5. Схема V-критика.

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 этого прогнозируемой ситуации $\mathbf{S_{pr}}(t+1)$ блок Критик может сделать оценку ее качества $V_{\mathrm{pr}} = V(\mathbf{S_{pr}}(t+1)) = V\,(\mathbf{S_{pr}}\,(\mathbf{S}(t), \mathbf{A}(t)))$.

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 следующее состояние $\mathbf{S_{pr}}(t+1) = \mathbf{S_{pr}}\,(\mathbf{S}(t), \mathbf{A}(t))$. Критик оценивает качество прогнозируемой ситуации $V_{\mathrm{pr}} = V(\mathbf{S_{pr}}(t+1))$. Действие $\mathbf{A}(t)$ выполняется, анимат получает награду $r(t)$. Оценивается ошибка временной разности:

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

Обучаются Критик:

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

и Контроллер:

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

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

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 [32,33].

Более полно теория адаптивных критиков и ее современное состояние характеризуется в работах [44,45].

2.4. Теория функциональных систем П.К. Анохина как концептуальная основа исследований адаптивного поведения

Для осмысления многообразия форм адаптивного поведения необходимо не только исследование конкретных моделей, но и разработка общих концепций и схем, позволяющих взглянуть сверху, "с высоты птичьего полета" на эти исследования.

Из одной таких концептуальных теорий может служить теория функциональных систем, предложенная и развитая в 1930-70 годах известным советским нейрофизиологом П.К. Анохиным [46,47].

Функциональная система по П.К. Анохину – схема управления, нацеленного на достижение полезных для организма результатов.

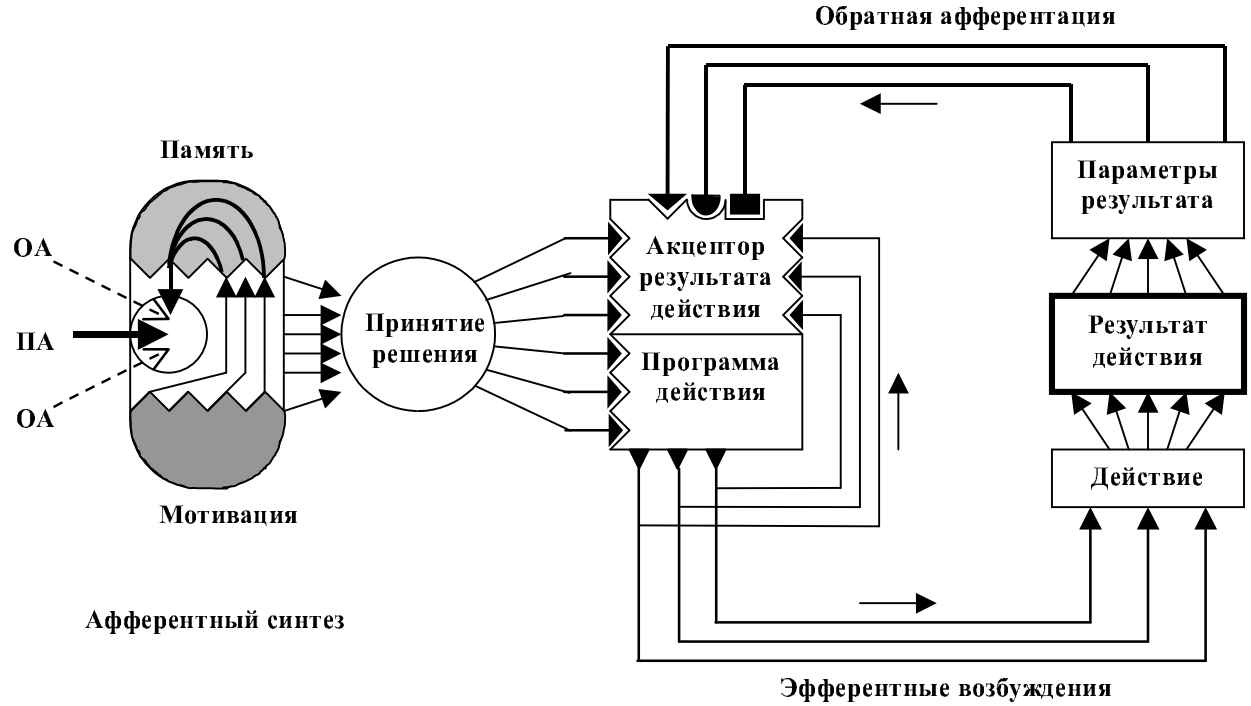

Рис. 6. Структура функциональной системы. ОА – обстановочная афферентация, ПА – пусковая афферентация.

Работа функциональной системы (рис. 6) может быть описана следующим образом.

Сначала происходит *афферентный синтез*, который включает в себя нейронные возбуждения, обусловленные 1) доминирующей мотивацией (понятие "мотивация" кратко обсуждается ниже), 2) обстановочной и пусковой афферентацией, 3) врожденной и приобретаемой памятью.

За афферентным синтезом следует *принятие решения*, при котором происходит уменьшение степеней свободы для эфферентного синтеза и выбор конкретного действия в соответствии с доминирующей мотивацией и с другими составляющими афферентного синтеза.

Затем следует формирование *акцептора результата действия*, т.е. прогноза результата. Прогноз включает в себя оценку параметров ожидаемого результата.

Эфферентный синтез – подготовка к выполнению действия. При эфферентном синтезе происходит генерация определенных нейронных возбуждений перед подачей команды на выполнение действия.

Все этапы достижения результата сопровождаются *обратной афферентацией*. Если параметры фактического результата отличаются от параметров акцептора результата действия, то действие прерывается и происходит новый афферентный синтез. В этом случае все операции повторяются, до тех пор, пока не будет достигнут конечный потребный результат.

Таким образом, функциональная система имеет циклическую (с обратными афферентными связями) саморегулирующуюся архитектонику.

Теория П.К. Анохина подразумевает *динамизм функциональных систем*. Для каждого конкретного поведенческого акта может быть сформирована своя функциональная система.

Функциональные системы формируются в процессе *системогенеза*. Теория системогенеза, которая исследует закономерности формирования функциональных систем в *эволюции*, *индивидуальном развитии* и *обучении* [48], может рассматриваться как отдельная ветвь теории функциональных систем. Отметим, что указанные составляющие системогенеза соответствуют составляющим формирования систем адаптивного поведения в трактовке AinimatLab [20] (см. раздел 2.2).

Каждая функциональная система ориентирована на достижение *конечного потребного результата*.

Необходимо подчеркнуть, что теория функциональных систем была разработана, в первую очередь, для интерпретации нейробиологических данных и зачастую сформулирована в очень интуитивных терминах. Поэтому, хотя она и хорошо известна, она не общепризнанна и практически не использовалась при разработке серьезных моделей адаптивного поведения. Можно сказать, что попытки формализации теории функциональных систем только начинаются [49-52]. Тем не менее, эта теория базируется на многочисленных биологических экспериментальных данных и представляет собой хорошую концептуальную основу для исследования широкого спектра проблем адаптивного поведения.

Отталкиваясь от теории П.К. Анохина, можно предложить общую кибернетическую схему управления целенаправленным адаптивным поведением естественного или искусственного организма (рис. 7). Здесь под организмом можно подразумевать как животное, так и робот или социально-экономическую систему: промышленную фирму, государство, человечество.

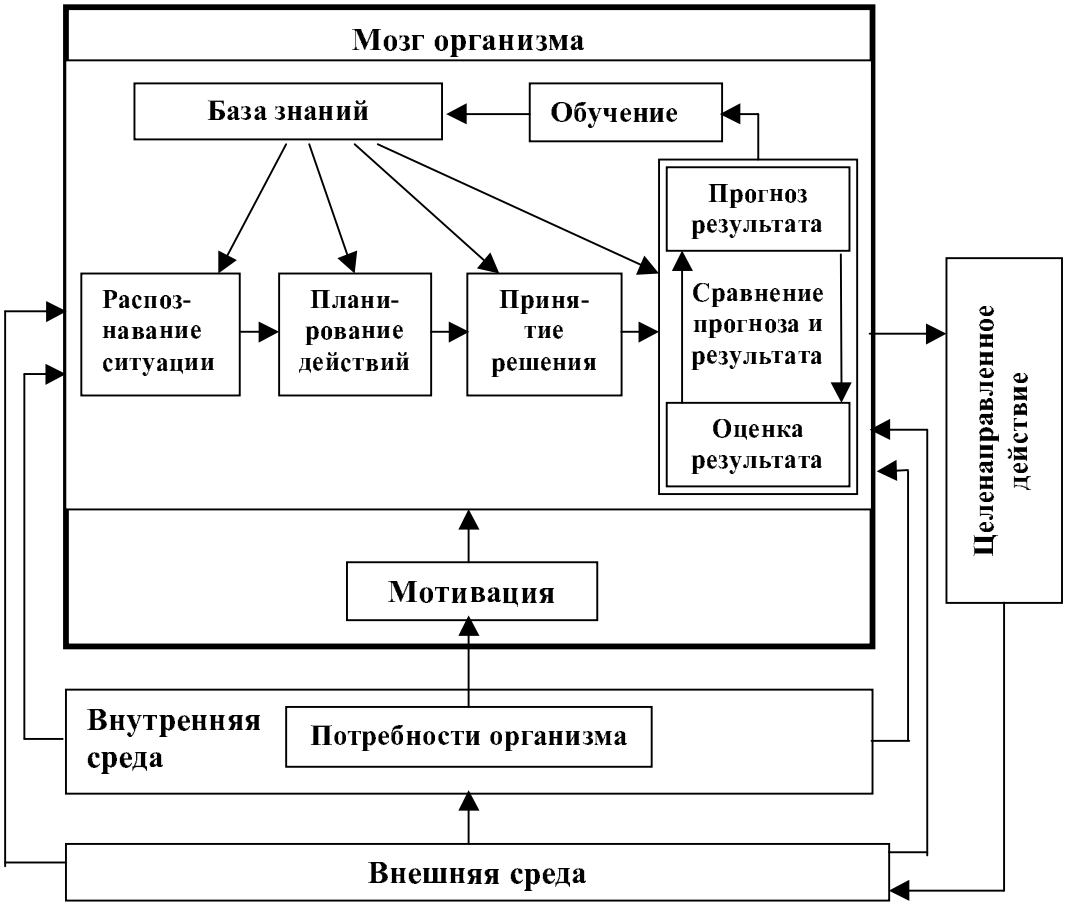

Рис. 7. Общая схема управления целенаправленным адаптивным поведением (в духе П.К. Анохина).

Теория функциональных систем обеспечивает концептуальную основу для построения конкретных моделей и для разработки общих проектов, направленных на разработку широкого спектра моделей адаптивного поведения. Далее в разделе 2.5 кратко характеризуются конкретные модели эволюционного возникновения целенаправленного адаптивного поведения, а разделе 2.6 излагается основанный на теории функциональных систем проект «Мозг анимата», который ориентирован на формирование общей платформы для исследования и моделирования широкого круга проблем адаптивного поведения.

Важное понятие функциональной системы – *мотивация*. Роль мотивации состоит в формировании цели и поддержке целенаправленных форм поведения. Мотивация может рассматриваться как активная движущая сила, которая стимулирует нахождение такого решения, которое адекватно потребностям организма в рассматриваемой ситуации. И имеет смысл провести моделирование эволюционного возникновения *целенаправленного* адаптивного поведения и анализ роли мотиваций в формировании целенаправленного поведения. Также следует отметить, что целенаправленность могла возникнуть на очень ранних стадиях эволюции, до появления каких-либо форм индивидуально приобретаемой памяти [53], поэтому, следуя пути, пройденному эволюцией, разумно начать с анализа этого свойства. Кроме того, свойство целенаправленности важно само по себе – это существенная особенность поведения *именно живых существ*.

Модели эволюционного возникновения целенаправленного адаптивного поведения были построены и исследованы в работах [36-38]. Основные результаты этого моделирования излагаются в следующем разделе.

2.5. Модели эволюционного возникновения целенаправленного адаптивного поведения

2.5.1. Модель «Кузнечик». Роль мотиваций в формировании адаптивного поведения [36,37]

В данной модели исследовался возможный механизм эволюционного возникновения целенаправленного поведения, обусловленного мотивациями.

Основные предположения модели состоят в следующем:

– Имеется популяция агентов (искусственных организмов), имеющих естественные потребности: 1) *потребность энергии* и 2) *потребность размножения*.

– Популяция эволюционирует в одномерной клеточной среде (рис. 8), в клетках может эпизодически вырастать трава (пища агентов). Каждый агент имеет *внутренний энергетический ресурс R* , который пополняется при съедании травы и расходуется при выполнении каких-либо действий. Уменьшение ресурса до нуля приводит к смерти агента. Агенты могут скрещиваться, рождая новых агентов.

– Потребности характеризуется количественно *мотивациями*. Если энергетический ресурс *R* агента уменьшается, то возрастает мотивация к пополнению энергетического ресурса (соответствующая потребности энергии) и уменьшается мотивация к размножению. При увеличении *R* мотивация к пополнению ресурса уменьшается, а мотивация к размножению растет.

– Поведение агента управляется его *нейронной сетью*. Сеть имеет один слой нейронов. На входы нейронов подаются сигналы, характеризующие внешнюю и внутреннюю среду агента, выходы нейронов определяют действия агента. Каждому возможному действию соответствует ровно один нейрон. В каждый такт времени совершается действие, соответствующее максимальному сигналу на выходе нейрона.

– Агенты "близорукие" – агент воспринимает состояние внешней среды только из трех клеток его поля зрения (рис. 8): той клетки, в которой агент находится, и двух соседних клеток.

– Агент может выполнять следующие *действия*: 1) быть в состоянии покоя ("отдыхать"), 2) двигаться, т.е. перемещаться на одну клетку вправо или влево, 3) прыгать через несколько клеток в случайную сторону, 4) есть (питаться), 5) скрещиваться. В силу способности агентов прыгать, мы называем их «кузнечиками».

– Нейронная сеть имеет специальные входы от мотиваций. Если имеется определенная мотивация, то поведение агента может меняться с тем, чтобы удовлетворить соответствующую потребность. Такое поведение можно рассматривать как *целенаправленное* (есть цель удовлетворить определенную потребность).
– Популяция агентов *эволюционирует*. Веса синапсов нейронной сети, управляющей поведением агента, составляют геном агента. Геном потомка формируется на основе геномов родителей при помощи рекомбинаций и мутаций.
– Мотивация к пополнению энергетического ресурса M_E и мотивация к размножению M_R определялись как простые функции энергетического ресурса агента R :

$$M_E = \max\{(R_0-R)/R_0\ ,\ 0\},\ \ M_R = \min\{R/R_1,\ 1\}\ ,$$

где R_0 , R_1 – параметры (обычно полагалось $R_0 = 2\ R_1$).

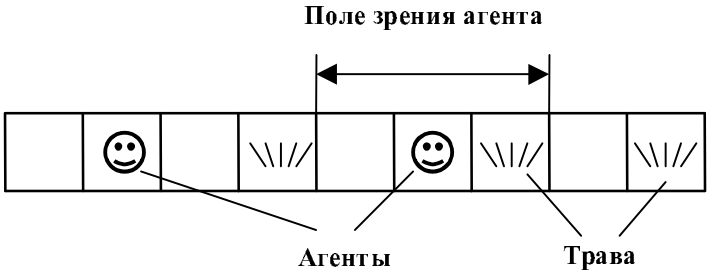

Рис. 8. Агенты в одномерной клеточной среде.

Модель исследовалась путем компьютерного моделирования эволюции популяции агентов. Нейронная сеть агентов исходной популяции определяла некоторые простые изначальные инстинкты, обеспечивающие питание и размножение агентов. Далее наблюдалось, как в процессе эволюции изменялись нейронная сеть агентов и определяемое ей поведение агентов.

Для того чтобы исследовать влияние мотиваций на поведение агентов, были проведены две серии компьютерных экспериментов. В первой серии моделировалась эволюция популяции агентов с "выключенными" мотивациями (входы нейронов от мотиваций были "задавлены"), во второй серии мотивации "работали" (так, как это изложено выше).

Основные результаты проведенного моделирования таковы:
– Мотивации играют важную роль в исследованных эволюционных процессах. А именно, если сравнить популяцию агентов без мотиваций с популяцией агентов с мотивациями, то, как показывают компьютерные эксперименты, эволюционный процесс приводит к тому, что вторая популяция (с мотивациями) имеет значительные селективные преимущества по сравнению с первой (без мотиваций). Этот вывод иллюстрируется рис. 9.
– Анализ нейронных сетей и поведения агентов демонстрирует, что управление поведением агента без мотиваций (рис. 10) можно рассматривать как набор простых инстинктов (несколько отличающихся от изначально заданных), а управление агентом с мотивациями (рис. 11) – как *иерархическую систему управления*, состоящую из двух уровней: уровня простых инстинктов и метауровня, обусловленного мотивациями. При этом иерархическая система управления обеспечивает более эффективное управление, чем одноуровневая система, в которой поведение определяется одними лишь простыми инстинктами. Переход от схемы управления без мотиваций (рис. 10) к схеме управления с мотивациями (рис. 11) подобен метасистемному переходу от простых рефлексов к сложному рефлексу в теории метасистемных переходов В.Ф. Турчина [54].

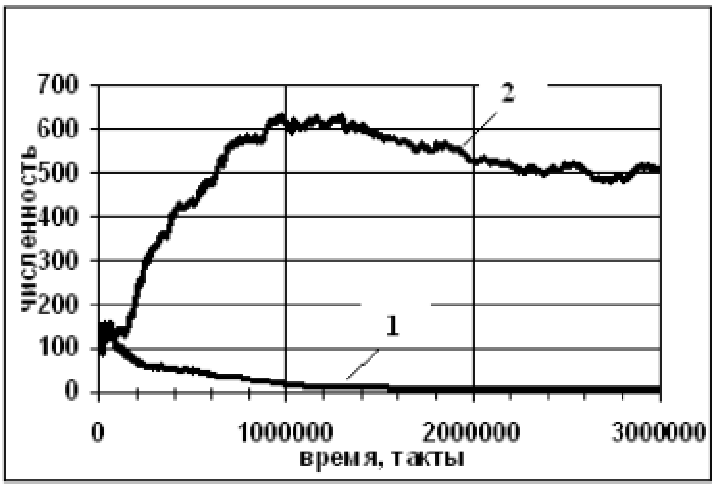

Рис. 9. Пример зависимостей численности популяции от времени для агентов без мотиваций (1) и с мотивациями (2). Видно, что популяция агентов с мотивациями имеет значительные селективные преимущества по сравнению с популяцией агентов без мотиваций.

Рис. 10. Схема управления агента без мотиваций. Поведение агента состоит из простых безусловных рефлексов, при котором выбор действия напрямую определяется текущим состоянием окружающей среды.

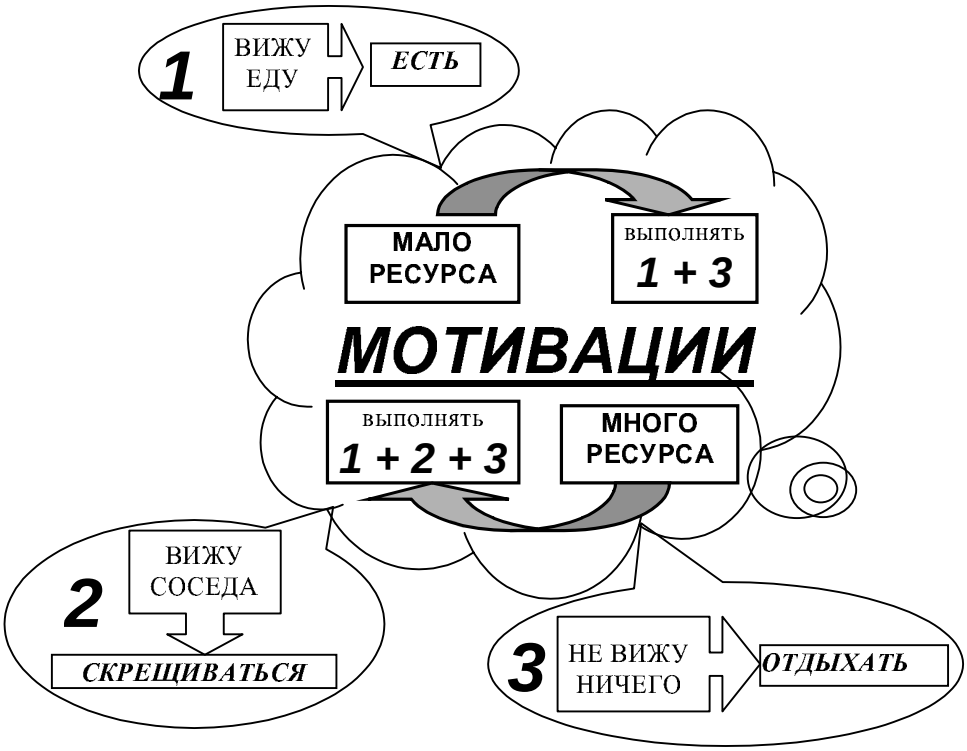

Рис. 11. Схема управления агента, обладающего мотивациями. Мотивации формируют новый уровень иерархии в системе управления агентами.

2.5.2. Развитие модели «Кузнечик» – возникновение естественной разветвленной иерархии целей

Изложенная модель была развита в работе М.С. Бурцева [38], в которой исследовалось поведение популяции агентов в двумерном мире (рис. 12). При этом дополнительно в модель были введены 1) возможность борьбы между агентами и 2) эволюционное изменение структуры нейронной сети, состоящей из рецепторов, эффекторов и связей между рецепторами и эффекторами.

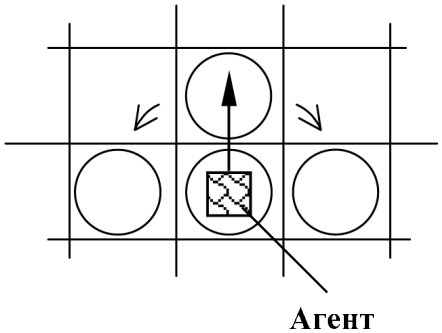

Рис. 12. Агент в двумерной клеточной среде. Агент ориентирован (стрелка показывает направление вперед), кружки – поле зрения агента. Действия агента: двигаться вперед, поворачиваться направо или налево, есть, размножаться, бороться с другими агентами. Система управления агента – однослойная нейронная сеть, оптимизируемая эволюционным методом.

Как и в предыдущей модели, для агентов исходной популяции задавалась некоторая минимальная система управления, обеспечивающая питание и размножение агентов. Поведение агентов начальной популяции (имеющих минимальный набор рецепторов и эффекторов) схематично представлено на рис. 13.

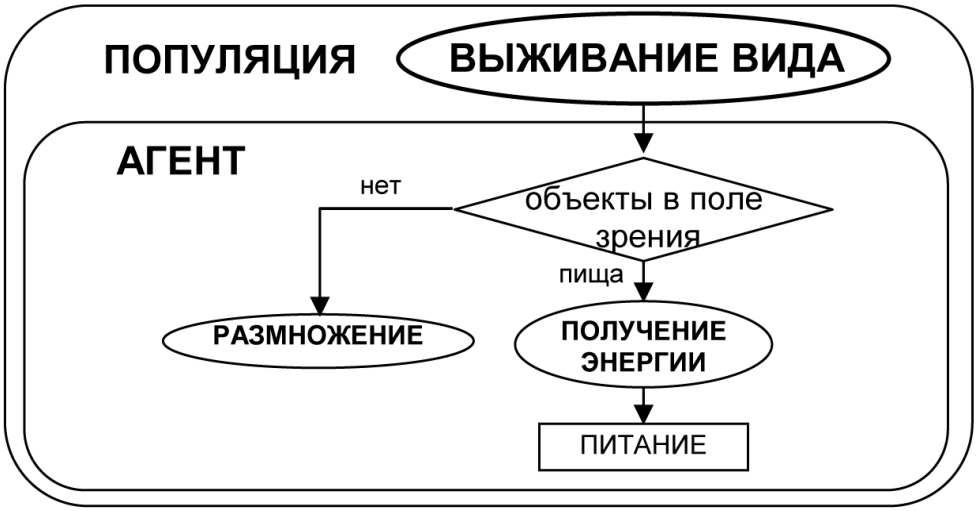

Рис. 13. Дерево условий для управления выбором подцелей агента начальной популяции.

В ходе эволюции поведение агентов структурируется. Стратегия агентов, сформированная в процессе эволюции, может быть представлена в виде схемы, показанной на рис. 14. Видно, что развивается достаточно сложное поведение, которое можно считать целенаправленным. Так первоначальный "инстинкт" агента, направленный на получение энергии, оптимизируется за счет появления еще одного уровня подцелей, направленных, соответственно: на собственно питание, на поиск пищи, борьбу.

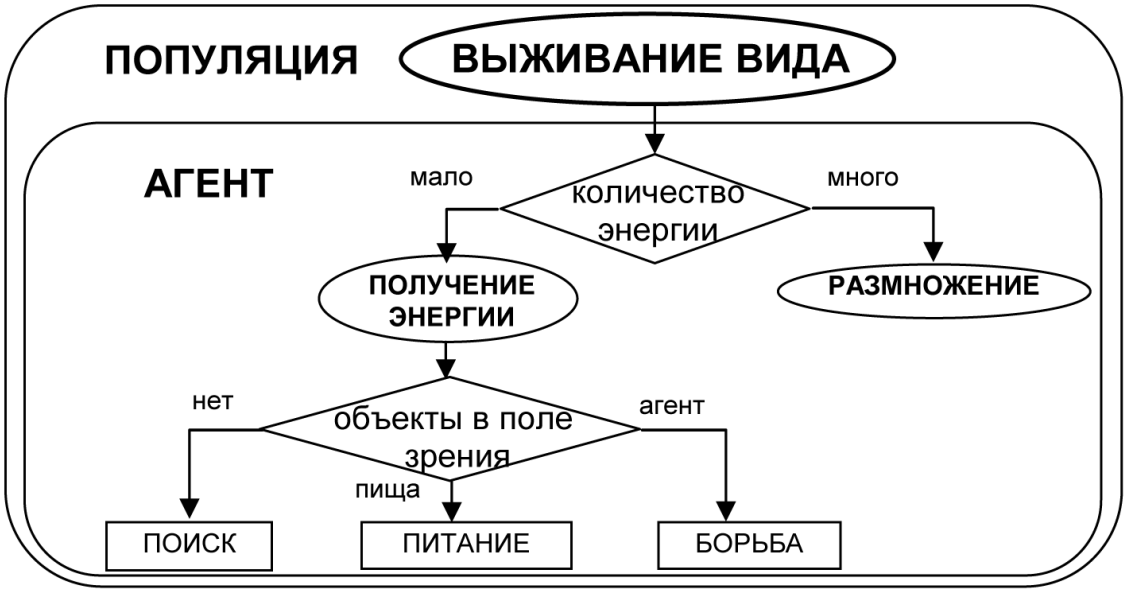

Рис. 14. Дерево условий для управления выбором подцелей, формирующееся в результате эволюции.

В целом моделирование, выполненное в [38], продемонстрировало, что в процессе исследованных эволюционных процессов возникает естественная иерархическая структура целей и подцелей.

2.6. Проект «Мозг анимата» [52] [1]

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

Отталкиваясь от теории функциональных систем П.К. Анохина (см. раздел 2.4), можно предложить достаточно общую «платформу» для систематического построения моделей адаптивного поведения. В работах [51,52] предложен проект «Мозг анимата», который как раз и нацелен на формирование общей схемы построения таких моделей. Кратко опишем данный проект. В работе [51] была предложена первая версия проекта, описывающая архитектуру системы управления анимата на основе нейросетевых блоков прогноза, обучаемых с помощью метода обратного распространения ошибки. Ниже излагается следующая версия архитектуры [52], основанная на нейросетевых адаптивных критиках (см. раздел 2.3).

Предполагается, что система управления аниматом имеет иерархическую архитектуру (рис.15). Базовым элементом системы управления является отдельная функциональная система (ФС).

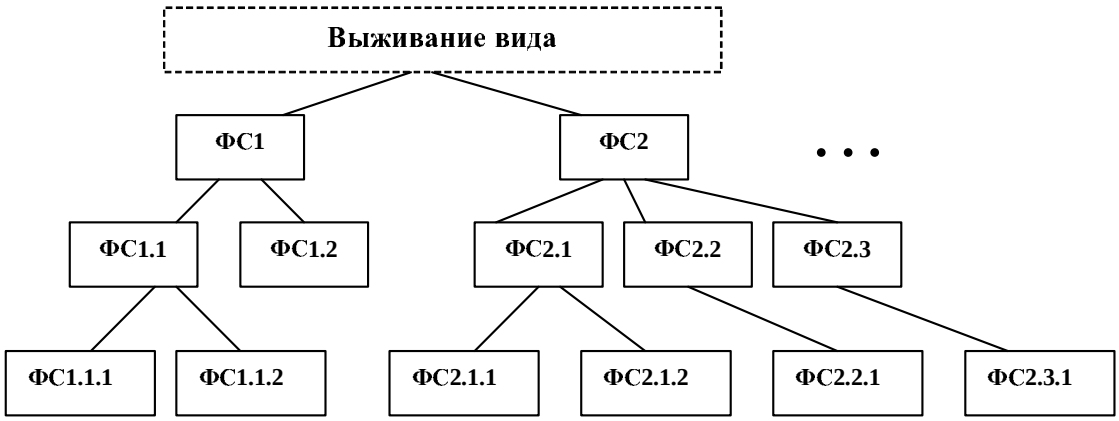

Рис. 15. Архитектура системы управления аниматом. ФС* – функциональная система.

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.

Самый верхний уровень соответствует выживанию вида (см. также схему иерархии управления на рис. 14). Этот уровень подразумеваемый, он не реализуется с помощью конкретной функциональной системы.

Предполагается, что система управления аниматом функционирует в дискретном времени. Также предполагается, что 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. Функциональная система использует одну из возможных схем адаптивных критиков, представленную ниже.

2.6.1. Схема адаптивного критика

Рассматриваемая схема адаптивного критика состоит из двух блоков: Модель и Критик (рис. 16). Предполагается, что Модель и Критик – многослойные персептроны, и что производные по весам нейронных сетей этих блоков могут быть вычислены обычным методом обратного распространения ошибки [55]. Предполагается, что адаптивный критик предназначен для выбора одного из нескольких действий. Например, при управлении движением действиями могут быть:

двигаться вперед, поворачивать вправо, поворачивать влево, стоять на месте. В каждый момент времени t адаптивный критик должен выбрать одно из возможных действий.

Цель адаптивного критика – максимизировать функцию суммарной награды $U(t)$ [19]:

$$U(t_j) = \sum_{k=0}^{\infty} \gamma^k r(t_{j+k}), \qquad (12)$$

где $r(t_j)$ – текущее подкрепление (награда, $r(t_j) > 0$ или наказание, $r(t_j) < 0$), полученное адаптивным критиком в данный момент времени t_j, γ – коэффициент забывания, $0 < \gamma < 1$. В общем случае разность t_{j+1} - t_j может зависеть от времени, но для простоты обозначений предполагается, что $\tau = t_{j+1} - t_j = \text{const}$.

Подчеркнем, что $U(t)$ есть алгебраическая сумма (с учетом коэффициента забывания) ожидаемых в будущем наград и наказаний. Значение $U(t)$ оценивается блоком Критик для каждой из возможных ситуаций $\mathbf{S}(t)$ величиной $V(\mathbf{S}(t))$. Эти оценки характеризуют качество той или ситуации и постепенно уточняются в процессе обучения. На основе этих оценок осуществляется выбор действий таким образом, чтобы максимизировать величину суммарной награды $U(t)$.

Принцип работы рассматриваемого адаптивного критика (рис.16) излагается ниже.

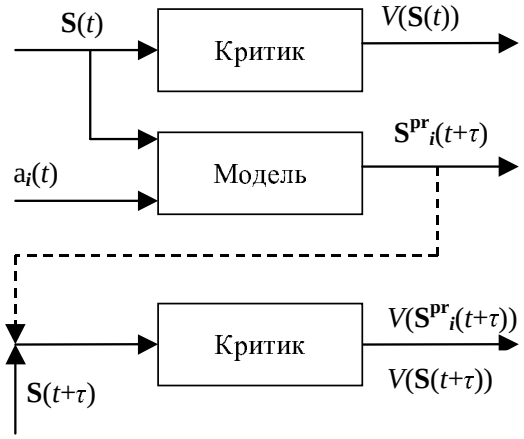

Рис. 16. Схема адаптивного критика, используемая в функциональной системе. Модель предсказывает следующую ситуацию $\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau)$ для всех возможных действий a_i, $i=1,2,\ldots, n_a$. Текущая ситуация $\mathbf{S}(t)$, прогнозы $\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau)$ и реальная следующая ситуация $\mathbf{S}(t+\tau)$ подаются на вход Критика (одна и та же нейронная сеть Критика показана в два последовательных момента времени), на выходе которого формируются оценки качества ситуаций $V(\mathbf{S}(t))$, $V(\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau))$ и $V(\mathbf{S}(t+\tau))$.

Модель имеет два типа входов: 1) входы, характеризующие текущую ситуацию $\mathbf{S}(t)$ (сигналы из внешней и внутренней среды анимата), и 2) входы, характеризующие действия. Предполагается, что каждое возможное действие a_i кодируется своей собственной комбинацией входов, и что число возможных действий невелико. Роль Модели – прогноз следующей ситуации для всех возможных действий a_i, $i=1,2,\ldots, n_a$.

Роль Критика – оценка качества ситуации $V(\mathbf{S})$ для текущей ситуации $\mathbf{S}(t)$, следующей ситуации $\mathbf{S}(t+\tau)$ и прогнозируемых ситуаций $\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau)$ для всех возможных действий. Величины V – оценки функции суммарной награды $U(t)$.

В каждый момент времени выполняются следующие операции:

1) Модель предсказывает следующую ситуацию $\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau)$ для всех возможных действий a_i, i=1,2,…, n_a .

2) Критик оценивает качество ситуации для текущей ситуации $V(t) = V(\mathbf{S}(t))$ и всех прогнозируемых ситуаций $V^{\mathbf{pr}}{}_i(t+\tau) = V(\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau))$.

1) Применяется ε -жадное правило [19], а именно, выбирается действие a_k следующим образом:

- с вероятностью 1 - ε выбирается действие с максимальным значением $V(\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau))$:
$k = \arg\ \max\ {}_i\{V(\mathbf{S}^{\mathbf{pr}}{}_i(t+\tau))\}$
- с вероятностью ε выбирается произвольное действие a_k, $0 < \varepsilon << 1$,

k – индекс выбираемого действия a_k .

4) Действие a_k выполняется.

5) Оценивается текущее подкрепление $r(t)$ и происходит переход к следующему моменту времени $t+\tau$. Наблюдается следующая ситуация $\mathbf{S}(t+\tau)$ и сравнивается с прогнозом $\mathbf{S}^{\mathbf{pr}}{}_k(t+\tau)$. Корректируются веса $\mathbf{W}_M$ нейронной сети Модели методом обратного распространения ошибки с целью минимизации ошибки прогноза:

$$\Delta\ \mathbf{W}_M = \alpha_M\ \mathrm{grad}_{\mathbf{W}M}(\mathbf{S}^{\mathbf{pr}}{}_k(t+\tau))^{\mathrm{T}}(\mathbf{S}(t+\tau) - \mathbf{S}^{\mathbf{pr}}{}_k(t+\tau))\ , \qquad (13)$$

где α_M – скорость обучения нейронной сети Модели.

6) Критик оценивает величину $V(\mathbf{S}(t+\tau))$. Считается ошибка временной разности [19]:

$$\delta(t)\ = r(t) + \gamma\ V\ (\mathbf{S}(t+\tau)) - V\ (\mathbf{S}(t))\ . \qquad (14)$$

Величина $\delta(t)$ характеризует ошибку в оценке $V(\mathbf{S}(t))$ – суммарной награды, которую можно получить, исходя из состояния $\mathbf{S}(t)$. Ошибка $\delta(t)$ рассчитывается с учетом текущей награды $r(t)$ и оценки суммарной награды $V(\mathbf{S}(t+\tau))$, которую можно получить, исходя из следующего состояния $\mathbf{S}(t+\tau)$.

7) Корректируются веса $\mathbf{W}_C$ нейронной сети Критика:

$$\Delta\ \mathbf{W}_C = \alpha_C\ \delta(t)\ \mathrm{grad}_{\mathbf{W}C}(V(t))\ , \qquad (15)$$

где α_C – скорость обучения нейронной сети Критика. Градиенты $\mathrm{grad}_{\mathbf{W}M}(\mathbf{S}^{\mathbf{pr}}{}_k(t+\tau))$ и $\mathrm{grad}_{\mathbf{WC}}(V(t))$ означают производные выходов нейронных сетей относительно соответствующих весов синапсов. Градиенты считаются так же, как в методе обратного распространения ошибки [55].

Обучение по формулам (14), (15) – минимизация ошибки $\delta(t)$ путем подстройки весов синапсов нейронной сети Критика градиентным методом.

Смысл обучения Модели по формуле (13) – уточнение прогнозов будущих ситуаций.

Смысл обучения Критика по формулам (14),(15) состоит в том, чтобы итеративно уточнять оценку качества ситуаций $V(\mathbf{S}(t))$ в соответствии с поступающими подкреплениями.

Изложенная схема адаптивного критика – ядро рассматриваемой функциональной системы. Множество функциональных систем формируют полную систему управления аниматом (рис. 15).

2.6.2. Функционирование системы управления аниматом

Детальная структура модели ФС представлена на рис. 17. В основу ФС положена изложенная выше схема адаптивного критика. Дополнительные свойства ФС по сравнению со схемой адаптивного критика таковы: 1) ФС дополнительно формирует команды субсистемам и посылает отчеты о результатах действий суперсистеме, и 2) сравнение между прогнозом $\mathbf{S}^{pr}_k(t+\tau)$ и результатом $\mathbf{S}(t+\tau)$ может быть отложено до момента $t+\tau$, когда поступит отчет от субсистем (детальней см. ниже). Связи данной ФС с супер/субсистемами показаны вертикальными жирными/пунктирными стрелками.

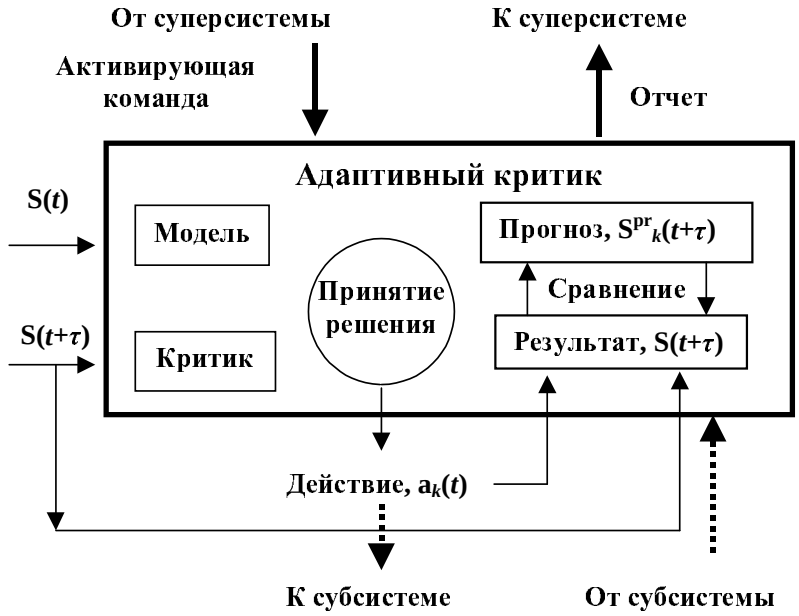

Рис. 17. Схема функциональной системы на основе адаптивного критика.

Предполагается следующая схема работы ФС в рамках функционирования всей системы управления аниматом. Данная ФС активизируется командой от суперсистемы. Модель и Критик функционируют так же, как описано выше в схеме работы адаптивного критика. В результате осуществляется выбор действия a_k. Дальнейшее зависит от вида действия a_k.

Если действие – команда для исполнительных элементов (сплошная стрелка вправо), то такое действие выполняется сразу; в этом случае $\tau = \tau_{min}$ – интервал между тактами времени минимален. Далее анимат получает подкрепление r из внешней или внутренней среды, и производится обучение в нейронных сетях адаптивного критика.

Другой тип действий – команды для субсистем (пунктирная стрелка вниз). Для такого действия подается команда активизации определенной субсистемы (выбор конкретной субсистемы определяется номером действия a_k). В этом случае сравнение прогноза и результата, оценка подкрепления r и обучение нейронных сетей откладывается до получения отчета от субсистемы, то есть до момента $t + \tau$, где $\tau > \tau_{min}$.

Обучение в обоих случаях осуществляется изложенным выше способом (раздел 2.6.1).

После выполнения всех этих действий ФС посылает отчет об окончании своей работы соответствующей суперсистеме.

Описанный способ работы ФС представляет собой обычный режим функционирования. Вводится также экстраординарный режим, который имеет место, если прогноз существенно отличается от фактического результата: $\| \mathbf{S}^{pr}_k(t_j) - \mathbf{S}(t_j) \| > \Delta > 0$, где $\|.\|$ обозначает некоторую норму, например,

евклидову. Предполагается, что в экстраординарном режиме величина ε (вероятность выбора случайного действия) в данной ФС и ее субсистемах резко возрастает, и поиск новых решений включает большую случайную компоненту. Этот поиск может сопровождаться случайным формированием и селекцией новых функциональных систем, аналогично селекции нейронных групп в теории нейродарвинизма Дж. Эдельмана [56]. Таким образом, обычный режим функционирования может рассматриваться как тонкая настройка системы управления аниматом, в то время как экстраординарный режим – это грубый поиск подходящего адаптивного поведения в чрезвычайных ситуациях.

Отметим, что в данную схему управления поведением анимата несложно включить процедуру прерывания верхними уровнями работы нижних уровней иерархии функциональных систем, с помощью специальных связей между ФС. Например, если в ФС1, отвечающую за безопасность, поступил сигнал, характеризующий серьезную опасность для жизни анимата, а анимат занимался поиском "пищи" в дереве решений, "возглавляемом" ФС2, ответственной за потребность питания, то ФС1 имеет право прервать работу ФС2 и дать команду на избежание опасности.

Память о старых навыках в нейронных сетях ФС может «портиться» при обучении новым навыкам, что соответствует известной дилемме пластичности-стабильности. Рассматриваемая архитектура системы управления аниматом позволяет включить естественным образом долговременную память о приобретенных навыках. Если некоторый тип поведения был хорошо апробирован, то соответствующая ему ФС может быть скопирована в долговременную память, а именно, в ФС, в которой величины ε and α_C , α_M равны нулю. Обе ФС – долгосрочная и краткосрочная, с долговременной и кратковременной памятью, соответственно – могут играть одну и ту же роль в общей архитектуре системы управления. Для надежных навыков долгосрочная ФС имеет приоритет по отношению к краткосрочной. Однако если прогнозы ситуаций $\mathbf{S}^{\mathbf{pr}}$, сделанные долгосрочной ФС, начинают отличаться от фактических **S**, то управление возвращается к краткосрочной ФС.

Итак, предложенная архитектура системы управления обеспечивает общий подход к моделированию адаптивным поведением анимата с естественными потребностями и соответствующими целями и подцелями. Сразу надо отметить, что использование адаптивных критиков в качестве функциональных систем – только один из возможных вариантов конструирования таких систем управления. Тем не менее, предложенная схема Мозга анимата позволяет уже сразу начинать работу по разработке конкретных моделей адаптивного поведения. По-видимому, одной из первых модельных реализаций могло бы быть воспроизведение адаптивного поведения агентов с иерархией целей и подцелей, описанного в разделе 2.5.2 (см. рис. 14).

Подчеркнем, что роль проекта «Мозг анимата» может быть глубокой и серьезной: этот проект может быть положен в основу базовых моделей «интеллектуальных» изобретений биологической эволюции (см. раздел 1 и рис. 2).

Проект «Мозг анимата» основан на нейросетевых адаптивных критиках, 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.

2.7. Модель эволюции популяции самообучающихся агентов на базе нейросетевых адаптивных критиков [57]

2.7.1. Описание модели

В данном разделе исследуется модель эволюции популяции самообучающихся автономных агентов и анализируется взаимодействие между обучением и эволюцией. Система управления

отдельного агента основана на нейросетевых адаптивных критиках [43,44] (см. также раздел 2.3). Модель отрабатывается на примере агента-брокера.

Схема агента-брокера. 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 [58]:

$$C(t+1) = C(t)\,\{1 + u(t+1)\,\Delta X(t+1) / X(t)\}\,[1 - J\,|u(t+1) - u(t)|], \qquad (16)$$

где $\Delta X(t+1) = X(t+1) - X(t)$ – текущее изменение курса акций, J – параметр, учитывающий расходы агента на покупку/продажу акций. Используется логарифмическая шкала для ресурса агента, $R(t) = \log C(t)$ [59]. Текущее подкрепление агента $r(t) = R(t+1) - R(t)$ равно:

$$r(t) = \log\{1 + u(t+1)\,\Delta X(t+1) / X(t)\} + \log[1 - J\,|u(t+1) - u(t)|]. \qquad (17)$$

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. 18). Цель адаптивного критика – максимизировать функцию полезности $U(t)$ [19]:

$$U(t) = \sum_{j=0}^{\infty} \gamma^j r(t+j), \qquad t = 1,2,..., \qquad (18)$$

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

Схема данного адаптивного критика подобна более общей схеме, предложенной в проекте «Мозг анимата» (раздел 2.6.1), однако имеет свою специфику, связанную с тем, что временной ряд $X(t)$ не зависит от действий агента.

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

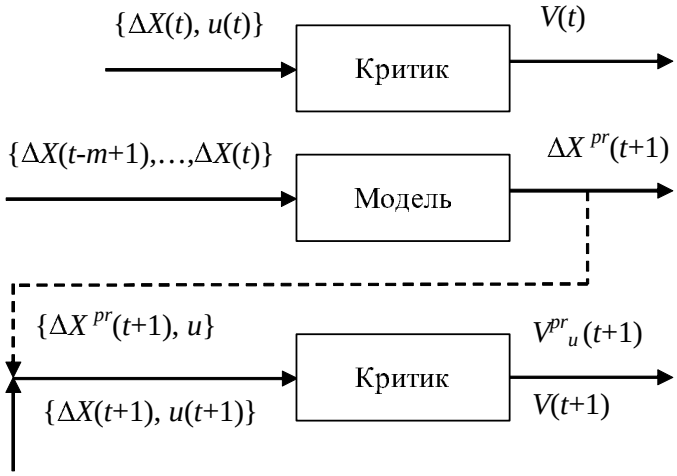

Рис. 18. Схема системы управления агента. НС Критика показана для двух последовательных тактов времени. Модель предназначена для прогнозирования изменения курса временного ряда. Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$ для текущей ситуации $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$, для ситуации в следующий такт времени $\mathbf{S}(t+1)$ и для предсказываемых ситуаций для обоих возможных действий $\mathbf{S}^{pr}{}_u(t+1) = \{\Delta X^{pr}(t+1), u\}$, $u = 0$ либо $u = 1$.

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

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

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 (18)) для агента, находящегося в рассматриваемой ситуации $\mathbf{S}$. Критик представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x^C} = \mathbf{S}(t) = \{\Delta X(t), u(t)\}, \quad y^C_j = \text{th}\,(\textstyle\sum_i w^C_{ij}\, x^C_i), \qquad V(t) = V(\mathbf{S}(t)) = \textstyle\sum_j v^C_j\, y^C_j\,,$$

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 [19]: 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 (17). 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:

$$\delta(t) = r(t) + \gamma\, V(t\text{+}1) - V(t)\,. \qquad (19)$$

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. Эволюционирующая популяция состоит 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 (17).

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

Начальные веса синапсов обоих 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+1)-ое поколение. Геномы потомков **G** отличаются от генома родителя небольшими мутациями.

Более конкретно предполагается, что в начале каждого нового (n_g+1)-го поколения для каждого агента его геном формируется следующим образом $G_i(n_g+1) = G_{best,\ i}\ (n_g) + \text{rand}_i$, $\mathbf{W}_0(n_g+1) = \mathbf{G}(n_g+1)$, где $\mathbf{G_{best}}(n_g)$ – геном лучшего агента предыдущего n_g-го поколения и rand_i – это $N(0, P^2_{mut})$, т.е., 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 $\mathbf{W} = \mathbf{W}_0 = \mathbf{G}$.

2.7.2. Результаты моделирования

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

1) синусоида:

$$X(t) = 0{,}5(1 + \sin(2\pi t/20)) + 1\ , \qquad (20)$$

2) стохастический временной ряд, использованный в [58]:

$$X(t) = \exp(p(t)/1200),\quad p(t) = p(t\text{-}1) + \beta(t\text{-}1) + k\,\lambda(t),\qquad \beta(t) = \alpha\beta(t\text{-}1) + \mu(t), \qquad (21)$$

где $\lambda(t)$ и $\mu(t)$ – два нормальных процесса с нулевым средним и единичной дисперсией, $\alpha = 0{,}9$, $k = 0{,}3$.

Некоторые параметры модели имели одно и то же значение для всех экспериментов: фактор забывания $\gamma = 0{,}9$; 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 $P_{mut} = 0{,}1$; расходы агента на покупку/продажу акций $J = 0$. 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 методом временной разности;
- 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 (20) показаны на рис. 19. 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. 19 видно резкое увеличение $R_{max}(n_g)$ после $n_g = 100$ и $n_g = 2000$ для соответствующих случаев). Результаты усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

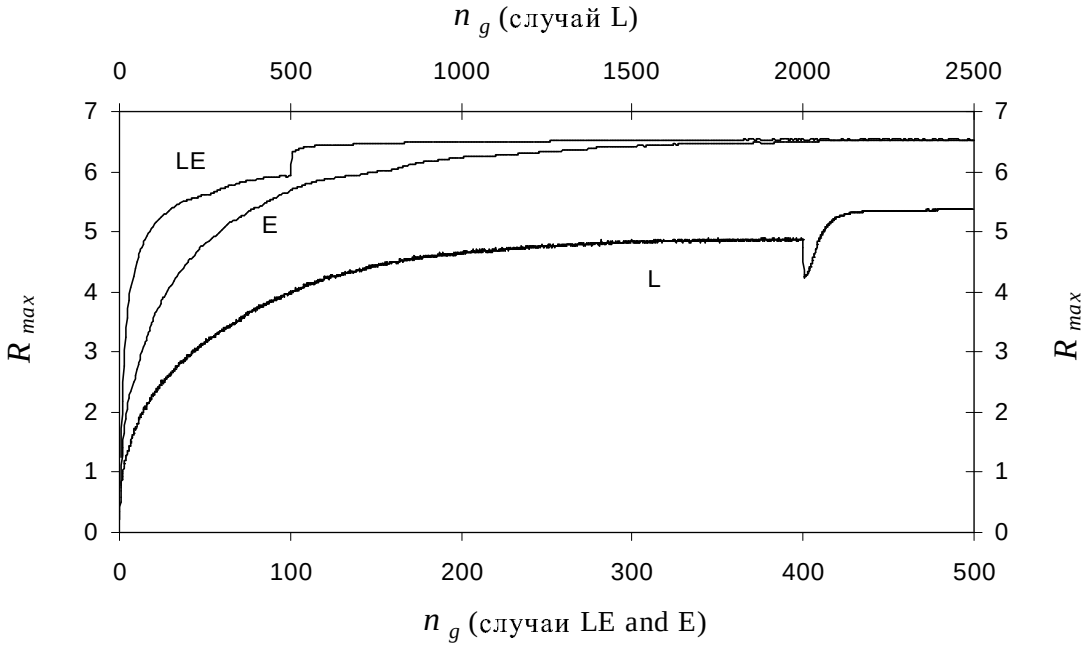

Рис. 19. Зависимости $R_{max}(n_g)$. Кривая LE соответствует случаю эволюции, объединенной с обучением, кривая E – случаю чистой эволюции, кривая L – случаю чистого обучения. Временная шкала для случаев LE и E (номер поколения n_g) представлена снизу, для случая L (индекс n_g) – сверху. Моделирование проведено для синусоиды, кривые усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

Рис. 19 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 (17) следует, что существует оптимальная стратегия поведения агента (затратами на покупку/продажу акций пренебрегалось, $J = 0$): вкладывать весь капитал в акции ($u(t+1) = 1$) при росте курса ($\Delta X(t+1) > 0$), вкладывать весь капитал в деньги ($u(t+1) = 0$) при падении курса ($\Delta X(t+1) < 0$). Анализ экспериментов, представленных на рис. 19, показал, что в случаях LE (обучение + эволюция), и E (чистая эволюция) такая оптимальная стратегия находится. Это соответствует асимптотическому значению ресурса $R_{max}(500) = 6{,}5$.

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 (21).

Итак, результаты, представленные на рис. 19, демонстрируют, что хотя обучение в настоящей модели и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения по сравнению со случаем чистой эволюции (см. графики LE и E на рис. 19).

Интересную особенность процесса поиска оптимального решения демонстрирует рис. 20. Этот рисунок показывает график $R_{max}(n_g)$ наряду со стандартным отклонением $\sigma(n_g)$ для случая LE. Значения $\sigma(n_g)$ характеризуют разброс значений $R_{max}(n_g)$ для различных реализаций моделируемых процессов. Рис. 20 показывает, что рост $R_{max}(n_g)$ сопровождается и ростом разброса значений R_{max} : кривая $\sigma(n_g)$ имеет максимум в области быстрого роста $R_{max}(n_g)$. Эта особенность имеет общий характер: кривые $\sigma(n_g)$ имеют аналогичные максимумы и для случаев L и E.

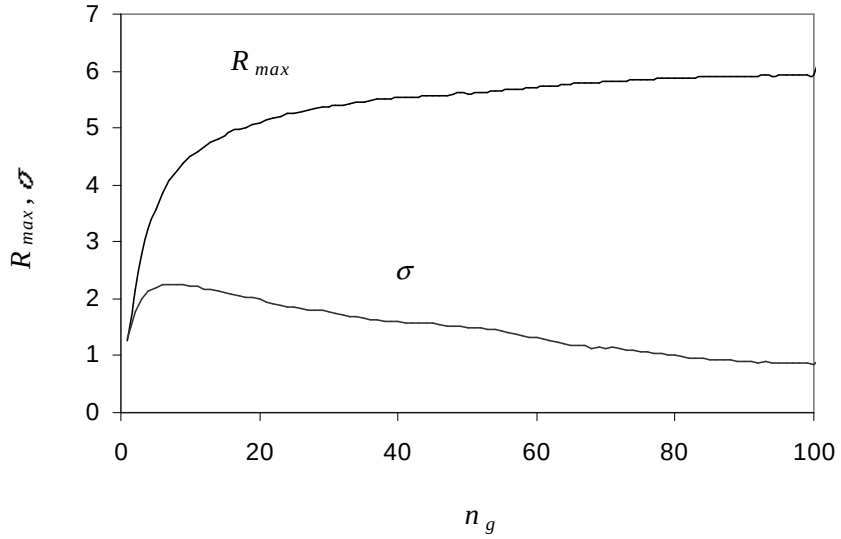

Рис. 20. Зависимости максимального по популяции ресурса $R_{max}(n_g)$, приобретаемого в течение поколения, и стандартного отклонения $\sigma(n_g)$ величины $R_{max}(n_g)$ от номера поколения n_g . Случай LE. Значения $\sigma(n_g)$ характеризуют разброс значений $R_{max}(n_g)$ для различных реализаций моделируемых процессов. Моделирование проведено для синусоиды, кривые усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

Стоит отметить, что эта особенность – увеличение числа случайных вариантов возможных решений на активной стадии поиска – сходна с явлением генерализации при выработке условного рефлекса [60]. При выработке условного рефлекса на стадии генерализации также происходит интенсификация случайной поисковой активности: реакция возникает не только на условный стимул, но на различные подобные ему (дифференцировочные) раздражители. И только затем происходит специализация, при которой реакция на дифференцировочные стимулы постепенно ослабевает и сохраняется только реакция на условный стимул. При выработке условного рефлекса зависимость условной реакции от числа экспериментов подобна кривой $R_{max}(n_g)$ на рис. 20. Увеличение интенсивности случайного поиска при генерализации сходно с ростом величины $\sigma(n_g)$ в области быстрого увеличения величины $R_{max}(n_g)$.

Особенности обучения (чистое обучение без эволюции). Рис. 19 демонстрирует, что рассмотренная простая форма обучения при данной структуре НС (см. п.2.7.1) несовершенна, так как она может привести лишь к «субоптимальной» стратегии поведения, даже если обучение происходит в течение большого числа поколений. Асимптотическое значение R_{max} для синусоиды составляет только R_{max}= 5,4 (см. кривую L на рис. 19), что значительно меньше асимптотического значения R_{max}= 6,5, соответствующего оптимальной стратегии (кривые LE и E на рис. 19). Это обусловлено тем, что чистое обучение способно найти лишь субоптимальную» стратегию: агент покупает акции, когда их цена растет или слегка падает и продает акции, когда их цена падает значительно. Такое поведение агента для синусоидального и стохастического временного ряда показано на рис. 21 и 22, соответственно.

Таким образом, в случае чистого обучения агент явно предпочитает хранить свой капитал в акциях, а не в деньгах.

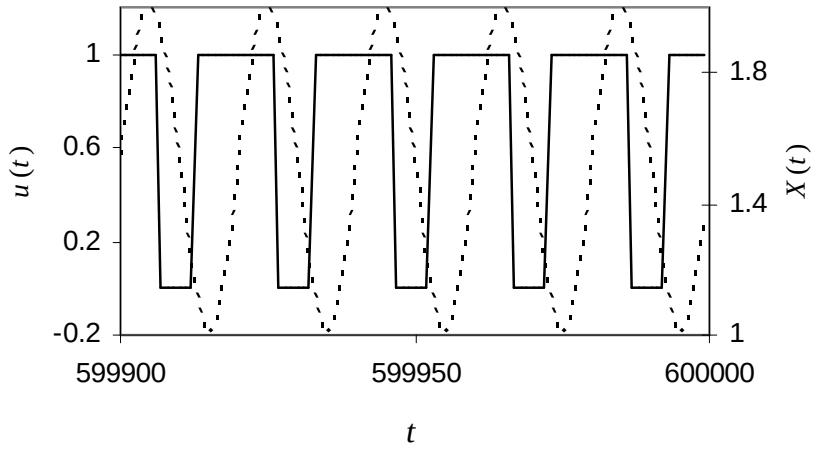

Рис. 21. Динамика поведения обучающегося агента для синусоиды (20). Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. Случай чистого обучения.

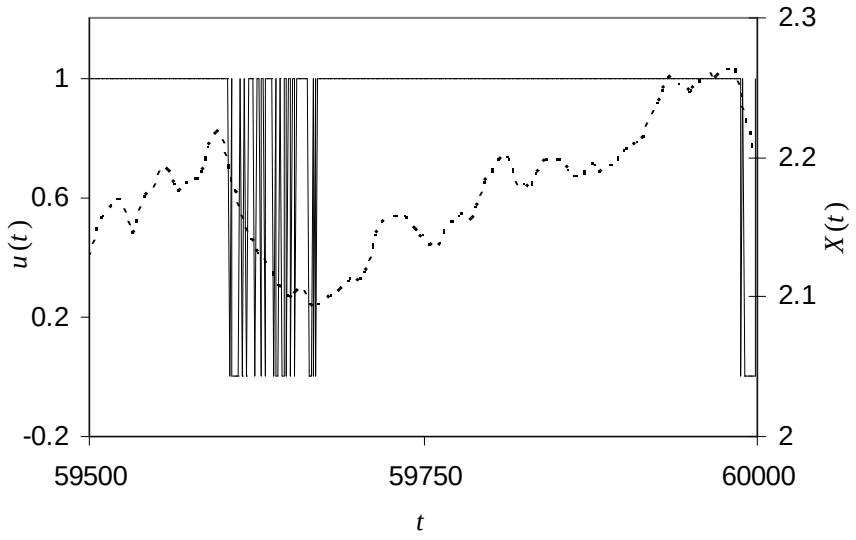

Рис. 22. Динамика поведения обучающегося агента для стохастического ряда (21). Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. Случай чистого обучения.

Взаимодействие между обучением и эволюцией. Эффект Балдвина. Как показано на рис. 19 (кривая E) для синусоидального временного ряда чистая эволюция способна найти оптимальную стратегию во всех экспериментах. В случае стохастического временного ряда оптимальная стратегия также обнаруживалась в экспериментах с чистой эволюцией, но лишь в некоторых расчетах. Например, при $N_g = 300$ и $T = 200$ эволюция смогла найти оптимальную стратегию в восьми из 10 экспериментов. Пример найденной таким образом оптимальной стратегии представлен на рис. 23.

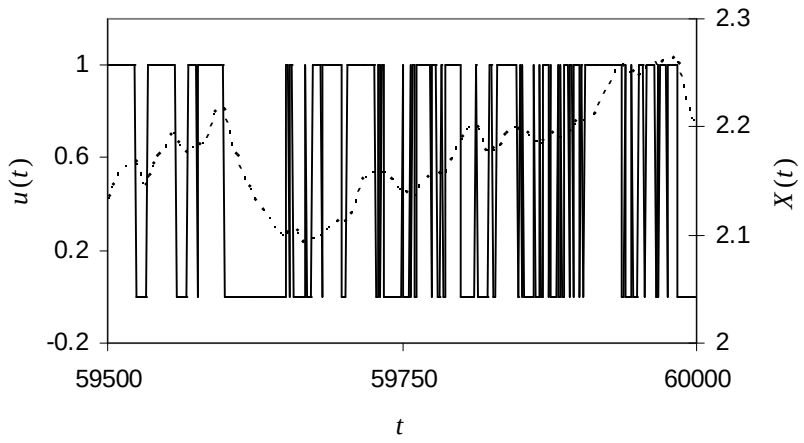

Рис. 23. Стратегия поведения лучшего агента в популяции. Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. $n = 10$, $T = 200$. Случай чистой эволюции. Стратегия поведения агента практически оптимальна: агент покупает/продает акции при росте/падении курса акций.

Рис. 19 также 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 Балдвина: исходно приобретаемый навык в течение ряда поколений становился наследуемым [61-63]. Этот эффект наблюдался в ряде экспериментов, один из которых представлен на рис. 24.

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 (20). Рис. 24 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. Таким образом, рис. 24 показывает, что стратегия, изначально приобретаемая посредством обучения, становится наследуемой (эффект Балдвина).

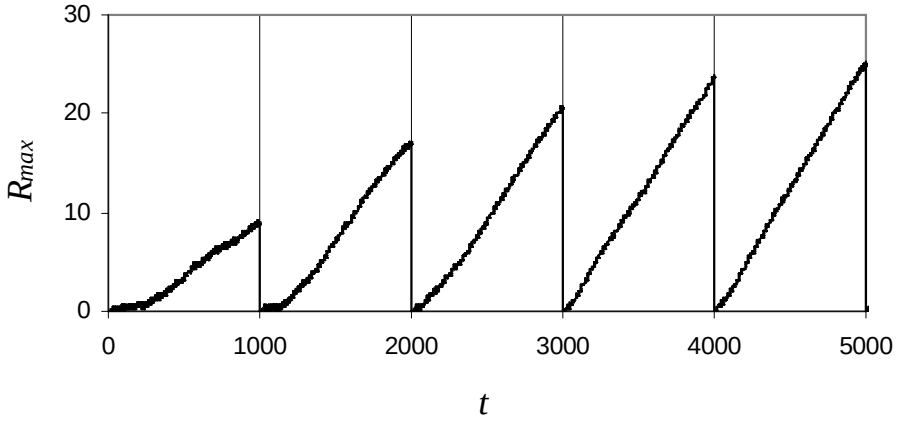

Рис. 24. Зависимость ресурса лучшего в популяции агента R_{max} от времени t для первых пяти поколений. Случай LE (эволюция, объединенная с обучением); размер популяции n = 10, длительность поколения T = 1000. Моменты смены поколений показаны вертикальными линиями. Для первых двух поколений есть явная 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 Балдвина 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 каждого агента включает в себя нейронную сеть Модели, предназначенную для предсказания изменения значения ΔX (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. Например, рис. 25 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 ΔX. Однако, предсказанные значения $\Delta X^{pr}(t+1)$ отличаются примерно в 25 раз от значений $\Delta X(t+1)$.

На рис. 26 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 подобна стратегии, представленной на рис. 23.

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 действий (покупать или продавать), приводящий к награде. А схема обучения предусматривает довольно сложную процедуру прогноза ситуации S, оценки качества прогнозируемых ситуаций, итеративного формирования оценок качества ситуаций V(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.

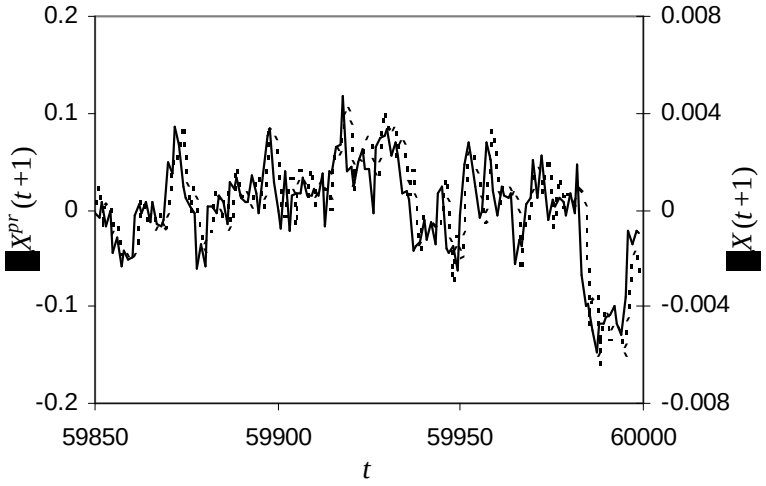

Рис. 25. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай чистой эволюции. n = 10, T = 200. Хотя обе кривые имеют сходную форму, по величине ΔX^{pr} (t+1) и ΔX (t+1) радикально различаются.

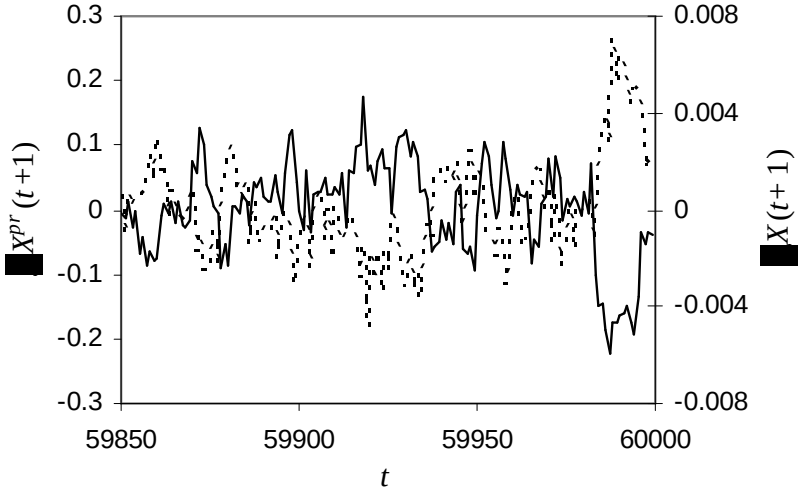

Рис. 26. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай эволюции, объединенной с обучением. n = 10, T = 200. Кривые ΔX^{pr} (t+1) и ΔX (t+1) различаются как по величине, так и знаком.

Сравнение с поведением простейших животных. Исследуемые агенты имеют две поведенческие тактики (продавать или покупать акции) и выбирают действия, переключаясь между этими тактиками. Можно сопоставить особенности этого поведения с переключением между двумя тактиками при поисковом поведении простейших животных. Например, некоторые виды личинок ручейников используют аналогичные тактики [30,64]. Личинки живут на речном дне и носят на себе «домик» – трубку из песка и других частиц, которые они собирают на дне

водоемов. Личинки строят свои домики из твердых частичек разной величины. Они могут использовать маленькие или большие песчинки [64]. Большие песчинки распределены случайно, но обычно встречаются группами. Используя большие песчинки, личинка может построить домик гораздо быстрее и эффективнее, чем используя маленькие, и, естественно, предпочитает использовать большие частицы. Личинка использует две тактики: 1) тестирование частиц вокруг себя и использование выбранных частиц, 2) поиск нового места для сбора частиц. Исследование поведения личинок обнаруживает инерцию в переключении с первой тактики на вторую [30,64]. Если личинка находит большую частицу, она продолжает тестировать частицы, пока не найдет несколько маленьких, и только после нескольких неудачных попыток найти новую большую частицу, переходит ко второй тактике. Во время поиска нового места личинка время от времени тестирует частицы, которые попадаются на ее пути. Она может переключиться со второй тактики на первую, если найдет большую частицу; при этом переключении также может проявляться инерция. Таким образом, переключение между тактиками имеет характер случайного поиска с явным эффектом инерции. Процесс инерционного переключения позволяет животному использовать только общие крупномасштабные свойства окружающего мира и игнорировать мелкие случайные детали.

В компьютерных экспериментах поведение агента-брокера, подобное поведению животных с инерционным переключением между двумя тактиками, наблюдалось, когда система управления агента оптимизировалась с помощью чистой эволюции при достаточно большой численности популяции. То есть фактически происходила оптимизация методом случайного поиска в достаточно большой области возможных решений. Рис. 27 показывает фрагмент стратегии поведения агента, найденной на ранней стадии эволюции в большой популяции, n = 100. Эта стратегия агента подобна описанному выше поведению животных с инерционным переключением между двумя тактиками. Стратегия переключения между u = 0 и u = 1 представляет собой реакцию только на общие изменения в окружающей среде (агент игнорирует мелкие флуктуации в изменении курса акций). Кроме того, переключение явно обладает свойством инерционности.

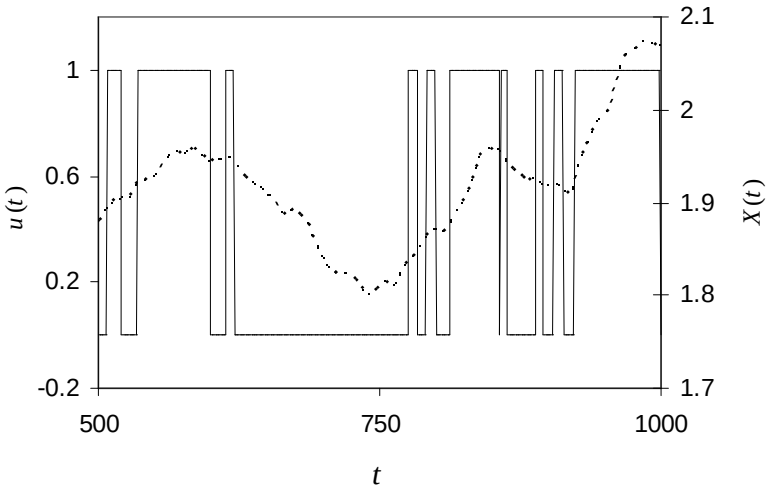

Рис. 27. Стратегия поведения лучшего агента в популяции. Действия агента характеризуются величиной $u(t)$ (сплошная линия): при u = 0 весь капитал переведен в деньги, при u = 1 весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. n = 100, T = 200. Стратегия агента подобна поведению животных с инерционным переключением между двумя тактиками. Агент игнорирует мелкие флуктуации динамики курса акций, переключение между при выборе действия u = 0 и выборе действия u = 1 обладает свойством инерционности.

2.7.3. Выводы по модели эволюции самообучающихся агентов

В рассмотренной модели оптимальная стратегия обеспечивается в случаях чистой эволюции или комбинации обучения и эволюции. Чистое обучение не обеспечивает нахождения оптимальной

стратегии. Тем не менее, хотя обучение и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения для случая комбинации обучения и эволюции по сравнению со случаем чистой эволюции.

При достаточно большой длительности жизни агентов часто наблюдалось и более явное влияние обучения на эволюционный процесс. В первых поколениях эволюционного процесса агенты явно обучались находить удачную стратегию поведения в течение своей жизни, а после смены ряда поколений такая стратегия была у агентов с самого рождения. То есть исходно приобретаемый навык в течение ряда поколений становился наследуемым, что можно интерпретировать как проявление известного эффекта Балдвина.

Моделирование продемонстрировало очень интересный феномен: нейронная сеть Модели может давать неверные предсказания, (Модель предсказывает правильно лишь форму изменений временного ряда, причем, возможно, с неправильным знаком), тем не менее, на основе этого неверного прогноза формируется оптимальная стратегия поведения агента. По-видимому, эта особенность работы системы управления агента обусловлена тем, что задача, которую «решает» эволюция в рассматриваемой модели значительно проще, чем та задача, которую решает обучение. Эволюции достаточно обеспечить выбор действий (покупать или продавать), приводящий к награде. А схема обучения предусматривает значительно более сложную процедуру прогноза ситуации, оценки качества прогнозируемых ситуаций, итеративного формирования оценок качества ситуаций 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» сложный механизм обучения.

Интересные особенности выявились при сравнении поведения модельного агента-брокера с поисковым поведением простых животных. Простые животные в поисковом поведении используют несколько тактик. Например, бабочка может чередовать 1) тактику двигаться в выбранном направлении и 2) искать новое направление движения, а ручейник чередует 1) тактику собрать частицы в удачном месте и 2) искать новое место для сбора частиц [30,31,64]. Переключение между тактиками у животных происходит с эффектом инерции, что позволяет животным игнорировать мелкие, случайные изменения в складывающихся ситуациях, и решать стоящие перед ними проблемы «стратегически», по крупному счету. При анализе поведения агента-брокера оказалось, что поведение, подобное поведению простых животных, находится самым простым путем – методом эволюционного поиска. Причем такое инерционное поведение формируется на ранних стадиях эволюции, единственное, что при этом требуется, чтобы была достаточно большая численность популяции.

Опыт работы с моделью показывает важность вопроса о том, какие системы управления автономных агентов являются эволюционно устойчивыми. Под эволюционной устойчивостью подразумевается свойство фенотипа (и соответствующего ему генотипа) становиться практически нечувствительным к мутациям. В частности, проведенное моделирование продемонстрировало, что сложные нейросетевые схемы обучения могут быть эволюционно нестабильны, если процесс обучения неустойчив относительно к возмущениям весов синапсов нейронных сетей.

Эволюционная неустойчивость работы рассмотренной схемы адаптивного критика показывает, что необходима определенная осторожность в выборе базовой модели функциональной системы (ФС) для проекта «Мозг анимата» (раздел 2.6). А, именно, хотя схема адаптивного критика, используемая в качестве основы ФС в изложенной в разделе 2.6 версии проекта, и моделирует прогноз результата действия и принятие решения на основе этого прогноза, что существенно для теории функциональных систем, тем не менее, имеет смысл рассмотреть и другие возможности для базовой модели ФС.

2.8. Выводы по моделям адаптивного поведения

Исследования адаптивного поведения – актуальное, содержательное и конструктивное направление, которое непосредственно связано с моделированием когнитивной эволюции, исследованием проблемы происхождения интеллекта. Также это направление исследований важно как биологически инспирированная научная основа разработок систем искусственного интеллекта. Это направление использует серьезные математические и компьютерные методы, и здесь построено множество интересных и содержательных моделей. Однако, результаты этих исследований пока достаточно скромные, в целом, результаты моделирования еще далеки от решения стратегически задач, поставленных при инициировании этого направления.

Один из значительных и достаточно неожиданных выводов этих исследований состоит в том, что часто нетривиальное поведение может быть сформировано простой системой управления [6]. Причем, таких систем управления, до которых сам конструктор системы управления может и не догадаться, а система управления формируется в процессе эволюционной самоорганизации, например, с помощью генетического алгоритма.

Определенная фрагментарность разработанных моделей показывает необходимость разработки общей «платформы» для систематизированного построения широкого спектра моделей адаптивного поведения. Такой платформой может стать изложенный выше проект «Мозг анимата».

3. Контуры программы будущих исследований

Анализ моделей адаптивного поведения показывает, что хотя проделана большая работа, ученые еще очень 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 возникновению интеллекта человека. То есть, есть огромная область чрезвычайно интересных исследований, которые только-только начинаются.

Предложим контуры программы будущих исследований проблемы происхождения интеллекта..

1. **Разработка схем и моделей адаптивного поведения анимата на базе проекта «Мозг анимата».**

   В разделе 2.6 изложен проект «Мозг анимата» [52], который предложен как общая «платформа» для систематизированного построения широкого спектра моделей адаптивного поведения. И реализация в моделях схем и конструкций Мозга анимата могла бы стать первым и важным шагом планируемых исследований.

   Воплощение в конкретные модели этих конструкций разумно начать с анализа целостного адаптивного поведения простых агентов, имеющих естественные потребности: питания, размножения, безопасности. Эволюционная схема формирования нейросетевой системы управления подобных агентов, обеспечивающей достаточно нетривиальную структуру целей и подцелей, была исследована М.С. Бурцевым [38] (раздел 2.5.2). Теперь было бы полезно промоделировать подобные системы управления в рамках конструкций Мозга анимата.

   Дальнейшая работа могла бы включать в себя анализ интеллектуальных изобретений биологической эволюции, таких как привыкание и условные рефлексы (рис. 2), на основе исследований проекта «Мозг анимата».

2. **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.

3. **Исследование процессов формирования причинной связи 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. При моделировании причинных связей было бы интересно «копнуть вглубь»: проанализировать процессы формирования таких связей на уровне отдельных нейронов и проследить схемы причинно-следственных связей от нейронного до поведенческого уровня.

   Естественный следующий шаг – переход от отдельных причинных связей к «базам знаний», к логическим выводам на основе уже сформировавшихся знаний.

4. **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 даже говорить, что такие выводы подобны выводам математика, доказывающего теоремы (раздел 1). И целесообразно разобраться в системах подобных выводов, понять, насколько адаптивна логика поведения животных и насколько она подобна нашей, человеческой логике. Возможно, здесь были бы полезны семантические сети, предложенные разработчиками искусственного интеллекта, и сопоставление процессов выводов на семантических сетях с «выводами» поведенческой логики животных.

5. **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.

Благодарности

Автор благодарен О.П. Мосалову за проведение части расчетов по модели агента-брокера (раздел 2.7), Д.В. Прохорову за многочисленные консультации по теории адаптивных критиков и В.А. Непомнящих за ряд обсуждений поведения биологических организмов.

ЛИТЕРАТУРА

1. Semantic Networks in Artificial Intelligence, Lehmann, Fritz, ed., Pergamon Press, Oxford, 1992.
2. Воронин Л.Г. Эволюция высшей нервной деятельности. М.: Наука. 1977. 128 с.
3. Meyer J.-A., Wilson S. W. (Eds) From animals to animats. Proceedings of the First International Conference on Simulation of Adaptive Behavior**.** The MIT Press: Cambridge, Massachusetts, London, England. 1990.

4. 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**.** The MIT Press: Cambridge, Cliff, Husbands, Meyer J.-A., Wilson S. W. (Eds) 1994, See also: http://animatlab.lip6.fr/index.en.html
5. Guillot A., Meyer J.-A. From SAB94 to SAB2000: What's new, Animat? // In Meyer et al. (Eds). From Animals to Animats 6. Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior. The MIT Press. 2000. See also: http://animatlab.lip6.fr/index.en.html
6. 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. См. также http://www.keldysh.ru/pages/BioCyber/RT/Nepomn.htm
7. Непомнящих В.А. Поиск общих принципов адаптивного поведения живых организмов и аниматов // Новости искусственного интеллекта. 2002. N. 2. С. 48-53.
8. 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. V. 26. N. 3. PP.381-395. See also: http://animatlab.lip6.fr/index.en.html
9. Wilson S.W. The animat path to AI // In: [3]. PP. 15-21.
10. Цетлин М.Л. Исследования по теории автоматов и моделирование биологических систем. – М.: Наука, 1969. 316 с.
11. Варшавский В.И., Поспелов Д.А. Оркестр играет без дирижера. М.: Наука, 1984.
12. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – "Животное" // Моделирование обучения и поведения. М.: Наука, 1975. С.152-171
13. Гаазе-Рапопорт М.Г., Поспелов Д.А. От амебы до робота: модели поведения. М.: Наука, 1987.
14. Holland J.H. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan Press, 1975 (1st edn). Boston, MA: MIT Press., 1992 (2nd edn).
15. Курейчик В.М. Генетические алгоритмы и их применение. Таганрог, ТРТУ, 2002.
16. Емельянов В.В., Курейчик В.М., Курейчик В.В. Теория и практика эволюционного моделирования. М.: Физматлит, 2003.
17. Редько В.Г. Эволюционная кибернетика. М.: Наука, 2001, 156 с.
18. Holland J.H., Holyoak K.J., Nisbett R.E., Thagard P. Induction: Processes of Inference, Learning, and Discovery. Cambridge, MA: MIT Press, 1986.
19. Sutton R., Barto A. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998. See also: http://www.cs.ualberta.ca/~sutton/book/the-book.html
20. Сайт AnimatLab: http://animatlab.lip6.fr/index.en.html
21. Сайт AI Laboratory of Zurich University: http://www.ifi.unizh.ch/groups/ailab/
22. Pfeifer R., Scheier C., Understanding Intelligence. MIT Press, 1999.
23. Сайт Laboratory of Artificial Life and Robotics: http://gral.ip.rm.cnr.it/
24. Nolfi S., Floreano D. Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines. Cambridge, MA: MIT Press/Bradford Books, 2000. 384 p.
25. Сайт MIT Computer Science and Artificial Intelligence Laboratory: http://www.csail.mit.edu/index.php
26. Brooks R.A. Cambrian Intelligence: The Early History of the New AI. MIT Press, 1999.
27. Сайт Neuroscience Institute: http://www.nsi.edu/
28. Krichmar J.L., Edelman G.M. Machine psychology: autonomous behavior, perceptual categorization and conditioning in a brain-based device // Cerebral Cortex, 2002, V. 12. PP. 818-830.
29. 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.
30. 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.
31. Nepomnyashchikh V.A., Podgornyj K.A. Emergence of adaptive searching rules from the dynamics of a simple nonlinear system // Adaptive Behavior. 2003. V.11. N.4. P.245-265.

32. Жданов А.А. Метод автономного адаптивного управления // Изв. РАН. Теория и системы управления. 1999. N. 5. С. 127-134.
33. Жданов А.А. О методе автономного адаптивного управления // VI Всероссийская научно-техническая конференция "Нейроинформатика-2004". Лекции по нейроинформатике. Часть 2. М.: МИФИ, 2004. С. 15-56.
34. Станкевич Л.А. Нейрологические средства систем управления интеллектуальных роботов // VI Всероссийская научно-техническая конференция "Нейроинформатика-2004". Лекции по нейроинформатике. Часть 2. М.: МИФИ, 2004. С. 57-110.
35. Самарин А.И. Модель адаптивного поведения мобильного робота, реализованная с использованием идей самоорганизации нейронных структур // IV Всероссийская научно-техническая конференция "Нейроинформатика-2002". Материалы дискуссии "Проблемы интеллектуального управления – общесистемные, эволюционные и нейросетевые аспекты". М.: МИФИ, 2003. С. 106-120.
36. Бурцев М.С., Гусарев Р.В., Редько В.Г. Модель эволюционного возникновения целенаправленного адаптивного поведения. 1. Случай двух потребностей // Препринт ИПМ РАН. 2000. N. 43. См. также http://www.keldysh.ru/pages/BioCyber/PrPrint/PrPrint.htm
37. Бурцев М.С., Гусарев Р.В., Редько В.Г. Исследование механизмов целенаправленного адаптивного управления // Изв. РАН. Теория и системы управления. 2002. N.6. С.55-62.
38. Бурцев М.С. Модель эволюционного возникновения целенаправленного адаптивного поведения. 2. Исследование развития иерархии целей // Препринт ИПМ РАН, 2002, N. 69.
39. Мосалов О.П., Редько В.Г., Непомнящих В.А. Модель поискового поведения анимата // Препринт ИПМ РАН, 2003, N. 19.
40. Мосалов О.П., Прохоров Д.В., Редько В.Г. Модели принятия решений на основе нейросетевых адаптивных критиков // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х т. М.: Физматлит, 2004. т.3. С. 1156-1163.
41. Klopf A. H. The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence. Hemisphere, Washington, 1982. 140 p.
42. Learning and Approximate Dynamic Programming: Scaling Up to the Real World (Edited by Jennie Si, Andrew Barto, Warren Powell, and Donald Wunsch), IEEE Press and John Wiley & Sons, 2004.
43. Редько В.Г., Прохоров Д.В. Нейросетевые адаптивные критики // Научная сессия МИФИ-2004. VI Всероссийская научно-техническая конференция "Нейроинформатика-2004". Сборник научных трудов. Часть 2. М.: МИФИ, 2004. С.77-84.
44. D.V. Prokhorov and D.C. Wunsch, “Adaptive critic designs,” *IEEE Trans. Neural Networks*, V. 8, no.5, pp. 997-1007, 1997.
45. D.V. Prokhorov, Backpropagation through time and derivative adaptive critics: a common framework for comparison. In [42]. See also: http://mywebpages.comcast.net/dvp/
46. Анохин П.К. Системные механизмы высшей нервной деятельности. М.: Наука, 1979. 453 с.
47. Анохин П.К. Принципиальные вопросы общей теории функциональных систем // Принципы системной организации функций. М.: Наука, 1973. С. 5-61.
48. Судаков К.В. (ред.). Теория системогенеза. М.: Горизонт, 1997.
49. Умрюхин Е.А. Механизмы мозга: информационная модель и оптимизация обучения. М.: Горизонт, 1999. 96 с.
50. Моделирование функциональных систем (под ред. Судакова К.В. и Викторова В.А.). М.: РАМН, РСМАН, 2000. 254 с.
51. 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 // Восьмая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. М.: Физматлит, 2002. Т.2. С.781-789.
52. Red'ko V.G., Prokhorov D.V., Burtsev M.S. Theory of functional systems, adaptive critics and neural networks // International Joint Conference on Neural Networks, Budapest, 2004. PP. 1787-1792.
53. Tsitolovsky L.E. A model of motivation with chaotic neuronal dynamics // Journ. of Biological Systems. 1997. V. 5. N.2. PP. 301-323.

54. Турчин В.Ф. Феномен науки. Кибернетический подход к эволюции. М.: Наука, 1993. 295с. (1-е изд). М.: ЭТС, 2000. 368 с. (2-е изд). См. также http://www.refal.ru/turchin/phenomenon/
55. Rumelhart D.E., Hinton G.E., Williams R.G. Learning representation by back-propagating error // Nature. 1986. V.323. N.6088. P. 533-536.
56. Edelman G. M. Neural Darwinism: The Theory of Neuronal Group Selection. Oxford: Oxford University Press, 1989.
57. Red'ko V.G., Mosalov O.P., Prokhorov D.V. A model of Baldwin effect in populations of self-learning Agents // International Joint Conference on Neural Networks, Montreal, 2005.
58. Prokhorov D., Puskorius G., Feldkamp L. Dynamical neural networks for control // In J. Kolen and S. Kremer (Eds.) A field guide to dynamical recurrent networks. NY: IEEE Press, 2001, PP. 257-289.
59. Moody J., Wu L., Liao Y., Saffel M. Performance function and reinforcement learning for trading systems and portfolios // Journal of Forecasting, 1998, vol.17, PP. 441-470.
60. Котляр Б.И., Шульговский В.В. Физиология центральной нервной системы. М.: Изд-во МГУ. 1979. 342 с.
61. Baldwin J.M. A new factor in evolution // American Naturalist, 1896, vol. 30, PP. 441-451.
62. Turney P., Whitley D., Anderson R. (Eds.). Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect // Special Issue of Evolutionary Computation on the Baldwin Effect, V.4, N.3, 1996.
63. Weber B.H., Depew D.J. (Eds.) Evolution and Learning: The Baldwin Effect Reconsidered. MA: MIT Press, 2003.
64. Nepomnyashchikh V.A. Selection behaviour in caddis fly larvae // In R. Pfeifer et al (Eds.) From Animals to Animats 5: Proceedings of the Fifth International Conference of the Society for Adaptive Behavior. Cambridge, MA: MIT Press, 1998, PP.155-160.