Как промоделировать сознание? *

В.Г. Редько
Институт оптико-нейронных технологий РАН
E-mail: vgredko@gmail.com

В конце названия этих тезисов умышленно поставлен знак вопроса, так как хотя тезисы и предлагают некоторые подходы к моделированию сознания, но это только подходы, только идеи, которые еще требуют осмысления и проработки.

Понятие «сознание» – многоплановое и, по-видимому, разные исследователи под этим термином подразумевают разные вещи. Здесь мы представим схемы компьютерного моделирования двух достаточно понятных аспектов сознания.

Первый аспект – формирование *поля внимания* [1,2] и внимательное, «сознательное» рассмотрение явления или процесса, помещенного в это поле.

Второй аспект – осознание своего собственного *субъективного Я*, и использование этого субъективного при восприятии внешнего мира [3,4].

1. Сознание и поле внимания

Со стороны кибернетики понятие «поле внимание» было введено в проекте «Животное» – нетривиальном проекте модели организации интеллектуального поведения, предложенном М.М. Бонгардом с сотр. в 1970-х годах [1]. В работе М.Н. Вайнцвайга и М.П. Поляковой [2] считается, что информационный процесс в мыслящей системе происходит сознательно тогда и только тогда, когда он проходит через поле внимания.

В нейробиологических работах А.М. Иваницкого также подчеркивается, что «…осознается только то, на что обращается внимание» [5]. Более того, в этих работах исследуются нейронные механизмы *возвратного возбуждения*, которые могут быть положены в основу моделирования процессов формирования полей внимания. Аналогичный механизм *повторного входа (re-entrance)* анализируется Дж. Эдельманом [6].

Предложим схему нейросетевого моделирования информационных процессов формирования поля внимания. Рассматриваем модельный организм; для определенности считаем, что организм соответствует уровню млекопитающего.

В основу модели положим понятие Хеббовского ансамбля [7] – множества нейронов, связанных между собой положительными связями. Такой ансамбль может рассматриваться как автоассоциативная память [8] – при возбуждении части нейронов за счет положительных связей возбуждается и весь ансамбль.

Пусть имеется множество ансамблей, в части ансамблей этого множества запоминаются элементарные сенсорные входные образы, в других ансамблях запоминаются обобщения сенсорных образов, представляющие собой понятия, формирующиеся в памяти организма. Имеются связи между ансамблями, кодирующими элементарные образы, и ансамблями, кодирующими обобщенные понятия, – эти связи обеспечивают естественное иерархическое соотношение между элементарными образами и понятиями, их обобщающими. Часть

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

ансамблей связана с эффекторами, обеспечивающими действия модельного организма. Также предполагаем, что на основе таких ансамблей и связей между ними формируются внутренние модели внешнего мира, позволяющие делать предсказания относительно событий во внешнем мире (в терминах теории функциональных систем предсказания соответствуют акцепторам результата действия [9]). В целом из ансамблей должна формироваться семантическая сеть, (аналогичная семантическим сетям в искусственном интеллекте [10]), обеспечивающая знания организма и управление поведением организма на основе знаний.

Каковы могут быть механизмы обучения такой системы управления? Предполагаем, что у организма есть жизненно важные потребности (размножения, питания, безопасности). Во время жизни организма он получает положительные или отрицательные подкрепления, связанные с потребностями. В соответствии с этими подкреплениями усиливаются или ослабляются связи между ансамблями. При сильной величине положительного или отрицательного подкрепления происходит модификация связей между активными ансамблями, а также происходит формирование новых ансамблей. Аналогично происходит формирование новых ансамблей и модификация связей между ансамблями, если оказались неверными предсказания о событиях во внешнем мире или о взаимодействии организма с внешним миром.

В процессе обучения активные ансамбли помещаются в поле внимания и при жизненно важных событиях осознаются – находятся в поле внимания достаточно длительное время, что обеспечивает необходимые для обучения модификации в нейронных сетях. Именно сознательное обращение внимание на события и образы при рассогласовании прогноза и результата или при жизненно важных событиях (сильное подкрепление или наказание) обеспечивает достаточно длительные модификации в нейронной сети, необходимые для пополнения или корректировки знаний. При этом знания накапливаются в нейронных сетях инкрементным образом, т.е. старые знания не исчезают – происходит только пополнение системы знаний о мире за счет добавления новых знаний [11].

Здесь мы не будем обсуждать детали механизма помещения событий в поле внимания – такой механизм может быть основан на совпадении во времени а) активности в нейронных ансамблях и б) наличия сигнала подкрепления/наказания или наличия сигнала рассогласования прогноза ситуации и реальной ситуации. Кроме того, при помещении событий в поле внимания важную роль может играть гиппокамп, обеспечивающий быстрый поиск адекватных ситуации знаний [5].

Для нас важно, что не так уж сложно разработать компьютерную модель сознательного формирования поля внимания: есть несколько простых механизмов обучения нейронных ансамблей [12] и несложно представить схемы акцентирования внимания на активных ансамблях.

Отметим, что очерченная выше схема формирования знаний на основе нейронных ансамблей аналогична теории ассоциативно-проективных нейросетей, которая разрабатывалась в конце 1980-х годов Э.М. Куссулем (Институт кибернетики, Киев) [13]. Также отметим, что в настоящее время достаточно нетривиальные компьютерные модели мозга, с особым акцентом на анализ роли гиппокампа, исследуются в Институте нейронаук Дж. Эдельмана [14].

Далее рассмотрим второй аспект: как можно промоделировать формирование субъективного Я. Здесь мы тоже постараемся показать, что нетрудно представить схему компьютерной модели эволюционного возникновения субъективного самосознания. Идея такой схемы

возникла в процессе дискуссии на сайте Рабочего совещания «О проблеме сознания», прошедшего во время конференции Нейроинфоматика-2006 [15].

2. Как промоделировать самосознание робота

Представим схему исследования, в котором мы можем промоделировать эволюционное происхождение субъективного Я.

Есть такое направление исследований – эволюционная роботика [16], в котором исследуется как путем эволюционного моделирования, т.е. в процессе эволюционной самоорганизации, формируются нейронные схемы управления роботов. Одно из интересных направлений в эволюционной роботике – исследование коллективного поведения роботов. При этом можно и не обязательно исследовать реальных роботов, а можно работать и с компьютерными моделями роботов, например, такими, каких исследует Л.А. Станкевич с сотр. при моделировании поведения команды виртуальных роботов-футболистов [17]. У таких роботов уже есть довольно сложная модульная «нервная система», архитектура которой включает три уровня: (1) физических действий, (2) индивидуального поведения, (3) координированного коллективного поведения. Отметим, что команда программистов под руководством Л.А. Станкевича, моделирующая виртуальных роботов, стала в 2004 году чемпионами мира в Симуляционной лиге футбола роботов.

Таким образом, есть серьезный задел исследований сложных блочных многоуровневых систем управления виртуальных и реальных роботов. И эти системы управления могут оптимизироваться эволюционным путем, путем эволюционной самоорганизации.

Теперь перейдем к главному моменту – как промоделировать возникновение субъективного Я. Пусть есть несколько популяций роботов (для определенности – виртуальных, существующих в форме компьютерных программ). И пусть эти популяции существуют в сложной среде, в которой есть питательный ресурс роботов, и те роботы, которые быстрее и эффективнее осваивают этот питательный ресурс, быстрее и размножаются. Популяции роботов могут конкурировать между собой: разные популяции существуют в одной и той же среде и могут бороться между собой за жизненный ресурс. «Нервная система» таких роботов – блочно-иерархическая и представляет собой развитие нервной системы роботов, аналогичных предложенным в [17].

Далее, пусть в нервной системе части роботов в одной из популяций возникает блок, ответственный за субъективное Я. Пусть он сначала возникает случайно, путем мутаций из других блоков. Так как нервная система роботов достаточно нетривиальная, то возникновение такого блока вполне вероятно.

Такой блок позволяет роботу с рассматриваемой нейронной сетью сказать: «Я – Робот». Этот блок позволяет данному роботу осознавать себя как личность и обеспечивает стремление стать важной личностью в своей популяции. Тогда такой робот может стать вожаком популяции и обеспечить единоначалие в принятии коллективных действий в данной популяции. Единоначалие и обусловленная им согласованность действий коллектива популяции, в свою очередь, обеспечивает селективное преимущество данной популяции перед другими популяциями, в которых у роботов нет блока, ответственного за субъективное. То есть, существование субъективного Я, сопровождаемого стремлением стать вождем, обеспечивает селективное преимущество и эволюционно устойчиво.

Понятно, что это только схема моделирования, которая вызывает множество вопросов. Но это схема вполне реального моделирования, показывающая, как эволюционно может

возникнуть и закрепиться субъективное Я.

В заключение еще раз подчеркнем, что выше очерчены только подходы к моделированию двух аспектов сознания, тем не менее, это реальные подходы к конкретному компьютерному моделированию информационных процессов в мозге, обеспечивающих сознательные процессы.

Литература

1. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – «Животное» // Моделирование обучения и поведения. – М.: Наука, 1975. С.152-171. Опубликовано также в книге: От моделей поведения к искусственному интеллекту. Серия "Науки об искусственном" (под ред. Редько В.Г.). М.: УРСС, 2006. С. 61-81.

2. Вайнцвайг М.Н., Полякова М.П. О моделировании мышления // От моделей поведения к искусственному интеллекту. Серия "Науки об искусственном" (под ред. Редько В.Г.). М.: УРСС, 2006. С. 280-286.

3. Дубровский Д.И. Сознание, мозг, искусственный интеллект // Статья на сайте Рабочего совещания "О проблеме сознания" конференции «Нейроинформатика-2006»: http://www.ni.iont.ru/NI06/WS/Dubrovsky.pdf

4. Базян А.С. Сознание, его формы и характеристики: соотношение сознания с памятью человека и животных // Статья на сайте Рабочего совещания "О проблеме сознания" конференции «Нейроинформатика-2006»: http://www.ni.iont.ru/NI06/WS/Bazyan.pdf

5. Иваницкий А.М. Проблема «Сознание и мозг» и искусственный интеллект // Научная сессия МИФИ-2006. VIII Всероссийская научно-техническая конференция "Нейроинформатика-2006": Лекции по нейроинформатике. М.: МИФИ, 2006. С. 74-87.

6. Edelman G.M. Group selection and phasic reentrant signaling: A theory of higher brain function. // In The Mindful Brain, Cambridge: MIT Press, 1978. PP.51-100.

7. Hebb D.O. The organization of behavior. A neuropsychological theory. N.Y.: Wiley & Sons, 1949. 355 p.

8. Редько В.Г. Эволюция, нейронные сети, интеллект. Модели и концепции эволюционной кибернетики. М.: КомКнига. (Изд-во УРСС, серия «Синергетика: от прошлого к будущему»), 2005. 224 с. Гл. 5.

9. Анохин П.К. Принципиальные вопросы общей теории функциональных систем // Принципы системной организации функций. – М.: Наука, 1973. Опубликовано также в книге: От моделей поведения к искусственному интеллекту. Серия "Науки об искусственном" (под ред. Редько В.Г.). М.: УРСС, 2006. С. 9-60.

10. Semantic Networks in Artificial Intelligence, Lehmann, Fritz, ed., Pergamon Press, Oxford, 1992.

11. Бурцев М.С. Классическая и эволюционная причинность в моделях обучения // 9-ая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. М.: Физматлит. 2004. Т.3. С. 1091-1098.

12. Фролов А.А., Муравьев И.П. Нейронные модели ассоциативной памяти. М.: Наука, 1987. 160 с

13. Куссуль Э.М. Ассоциативные нейроподобные структуры. Киев: Наукова думка, 1992. 144 с.

14. Krichmar J.L., Seth A.K., Nitz D.A., Fleischer J.G., Edelman G.M. Spatial Navigation and Causal Analysis in a Brain-Based Device Modeling Cortical–Hippocampal Interactions // Neuroinformatics, 2005, V.3, N.3. PP. 197–222.

15. Сайт Рабочего совещания "О проблеме сознания" конференции «Нейроинформатика-2006» (см. статьи А.С. Базяна и В.Г. Редько): http://www.ni.iont.ru/NI06/WS/Ws2006.htm

16. Nolfi S., Floreano D. Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines. Cambridge, MA: MIT Press/Bradford Books, 2000. 384 p.

17. Станкевич Л.А. Когнитивный подход к управлению гуманоидными роботами // От моделей поведения к искусственному интеллекту. Серия "Науки об искусственном" (под ред. Редько В.Г.). М.: УРСС, 2006. С. 386-443.