Подходы к разработке компьютерных моделей сознания*

Редько В.Г.
Научно-исследовательский институт системных исследований РАН, Москва

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

Первый аспект – эволюционное возникновение осознания простым животным схемы его организации.

Второй аспект – осознание своего собственного *субъективного Я* высокоорганизованными роботами, и использование этого субъективного Я при организации поведения.

Третий аспект – формирование *поля внимания* и внимательное, «сознательное» рассмотрение явления или процесса, помещенного в это поле.

1. Возможность моделирования эволюционного возникновения сознания

В работе Д.И. Дубровского[1] обсуждается вопрос: Почему субъективная реальность возникла в ходе биологической эволюции? Действительно, в ходе биологической эволюции произошли многоклеточные животные, управление целостным движением животного требовало осознания им всех частей организма. Хотя эта работа носит концептуальный характер, почти одновременно с ней появилось конкретное идейно близкое к ней моделирование по роботам[2]. При этом моделировании анализировалась динамика четырехногих роботов. В «сознании» робота формируется модель самого себя, т.е. внутренняя модель собственного тела в виде подвижного существа, состоящего из стержней и шарниров. При этом если происходила поломка робота (повреждалась одна из 4-х ног), то робот, как хромая собака, мог подстроить модель самого себя к поломке, и научиться перемещаться, прихрамывая. Существенно, что эта модель роботу была весьма полезна: робот, сознающий (конечно, примитивно) свое тело и перемещение тела во внешнем пространстве, двигался более эффективно, чем робот без такой модели. Таким образом, было продемонстрировано, что сознание субъективной реальности в форме модели движущегося во внешней среде организма полезно для этого организма. Понятно, что само такое моделирование должно быть развито, но, тем не менее, первый шаг сделан, и эффективность моделирования сознания простых организмов продемонстрирована.

Далее рассмотрим второй аспект: как можно промоделировать формирование субъективного Я в сообществе достаточно интеллектуальных роботов. Здесь мы постараемся показать, что нетрудно представить схему компьютерной модели эволюционного возникновения субъективного самосознания высокого уровня. Идея такой

* Работа выполнена при поддержке РФФИ (проект № 07-01-00180).

[1] Дубровский Д.И. Зачем субъективная реальность, или «Почему информационные процессы не идут в темноте?» (Ответ Д. Чалмерсу) // Дубровский Д.И. Сознание, мозг, искусственный интеллект. М.: Изд-во ИД Стратегия-Центр, 2007. С. 139-163.

[2] Bongard J., Zykov V., Lipson H. Resilient machines through continuous self-modeling // Science, 2006. V. 314. No 5802. PP. 1118-1121.

схемы возникла в процессе дискуссии на сайте Рабочего совещания «О проблеме сознания», прошедшего во время конференции Нейроинфоматика-2006[3].

2. Как промоделировать самосознание робота

Представим схему исследования, в котором можно промоделировать эволюционное происхождение субъективного Я в сообществе высокоорганизованных роботов.

Есть такое направление исследований – эволюционная роботика[4], 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 не обязательно исследовать реальных роботов, можно работать и с компьютерными моделями роботов, например, такими, которые исследуются Л.А. Станкевичем с сотрудниками при моделировании поведения команды виртуальных роботов-футболистов[5]. У таких роботов уже есть довольно сложная модульная «нервная система», управляющая поведением. Архитектура системы управления поведением робота включает три уровня: (1) физических действий робота, (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, имеется серьезный задел исследований сложных блочных многоуровневых систем управления виртуальных и реальных роботов. И эти системы управления могут быть оптимизированы эволюционным путем, путем эволюционной самоорганизации.

Теперь перейдем к главному вопросу данной схемы – как промоделировать возникновение субъективного Я. Пусть имеется 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 тем, которые разрабатывались Л.А. Станкевичем с сотрудниками.

Предположим, что 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. Этот блок позволяет роботу с рассматриваемой нейронной сетью сказать: «Я – Робот». Наличие блока дает возможность данному роботу осознавать себя как личность и обеспечивает стремление стать важной

---

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

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

[5] 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.

личностью в своей популяции. Тогда такой робот может стать вожаком популяции и обеспечить единоначалие в принятии коллективных решений 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 Я. Следовательно, 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 сообществе высокоорганизованных роботов.

3. Сознание и поле внимания

Со стороны кибернетики понятие «поле внимания» было введено в проекте «Животное» – нетривиальном проекте модели организации интеллектуального поведения, предложенном М.М. Бонгардом с сотрудниками в 1970-х годах[6]. Аналогично в работе М.Н. Вайнцвайга и М.П. Поляковой[7] 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»[8]. 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[9].

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[10] – множества нейронов, связанных между собой положительными связями. Такой ансамбль может рассматриваться как автоассоциативная память[11] – при возбуждении части нейронов за счет положительных связей возбуждается и весь ансамбль.

---

[6] 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.

[7] 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.

[8] 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.

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

[10] Hebb D.O. The Organization of Behavior. A Neuropsychological Theory. New York: Wiley and Sons, 1949. 355p.

[11] Редько В.Г. Эволюция, нейронные сети, интеллект. Модели и концепции эволюционной кибернетики. Серия «Синергетика: от прошлого к будущему». М.: УРСС, 2005. 224с. Гл. 5.

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 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[12]). В целом из ансамблей должна формироваться семантическая сеть (аналогичная семантическим сетям в искусственном интеллекте[13]), 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 внимания 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[14].

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.

---

[12] Анохин П.К. Принципиальные вопросы общей теории функциональных систем // Принципы системной организации функций. М.: Наука, 1973. С. 5-61. 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.

[13] Lehmann, F. (ed.). Semantic Networks in Artificial Intelligence, Oxford: Pergamon Press, 1992. 758p.

[14] 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.

Для нас важно, что не так уж сложно разработать компьютерную модель сознательного формирования поля внимания: есть несколько простых механизмов обучения нейронных ансамблей[15] 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)[16]. Также отметим, что в настоящее время достаточно нетривиальные компьютерные модели мозга, связанные с процессами в нейронных ансамблях, с особым акцентом на анализ роли гиппокампа, исследуются в Институте нейронаук Дж. Эдельмана[17].

В заключение подчеркнем, что выше изложены реальные подходы к конкретному компьютерному моделированию информационных процессов в мозге, обеспечивающих сознательные процессы.

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

[16] Куссуль Э.М. Ассоциативные нейроподобные структуры. Киев: Наукова думка, 1992. 144с.

[17] 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. No 3. PP. 197–222.