ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.8 Научная статья
DOI: 10.35330/1991-6639-2022-2-106-72-81
EDN: IHJBPX

ОСНОВНЫЕ МЕТОДЫ И ПОДХОДЫ К МОДЕЛИРОВАНИЮ ИСКУССТВЕННОГО СОЗНАНИЯ

И.А. ПШЕНОКОВА

Институт информатики и проблем регионального управления –
филиал Кабардино-Балкарского научного центра Российской академии наук
360000, Россия, Нальчик, ул. И. Арманд, 37-а

Аннотация. В статье приводится краткий анализ некоторых методов и подходов к моделированию искусственного сознания. Приводятся некоторые подходы к определению сознания в философии, психологии и нейробиологии. В частности, рассматриваются функциональные и нейробиологические модели сознания. Представлены некоторые подходы к моделированию искусственного сознания. Рассмотренные модели полностью удовлетворяют цели, для которой они были созданы, однако ни одна из них еще не показывает реальное создание личных предпочтений, приобретенных и обработанных через тело и эмоции агента, которые считаются основой для создания потенциального искусственного сознания. Эта область общего искусственного интеллекта активно развивается, и еще нет единой теории основополагающих принципов и методов создания интеллектуальных систем, обладающих сознанием, способных к пониманию своих действий и целей, а также самосознанию.

Ключевые слова: искусственный интеллект, сознание, искусственное сознание, робототехника, имитационное моделирование

Статья поступила в редакцию 11.02.2022 *Принята к публикации 10.03.2022*

Для цитирования. Пшенокова И.А. Основные методы и подходы к моделированию искусственного сознания // Известия Кабардино-Балкарского научного центра РАН. 2022. № 2 (106). С. 72–81. DOI: 10.35330/1991-6639-2022-2-106-72-81

Введение

В настоящее время значительно возрос интерес к возможности проектирования сознательных роботов, а значит, и искусственного сознания. Этот интерес частично основан на признании того, что сознание играет важную роль в принятии решений людей, и понимании того, что реализация модели сознания в интеллектуальных системах может помочь сделать роботов и искусственных агентов умнее. Но могут ли роботы быть сознательными и если да, то как проектировать такие машины? Для решения вопросов, характерных для искусственного сознания, рассмотрим сначала некоторые подходы к определению сознания.

В работе [1] сознание определено как представление субъекта о мире и о своем месте в нем, связанное со способностью дать отчет о своем внутреннем психическом опыте и необходимое для разумной организации совместной деятельности. В [2] сознание определяется как элемент высшей нервной деятельности человека. Обычно различают сознательную сущность, то есть сущность, которая является разумной, бодрствующей, имеет самосознание и субъективные качественные переживания и сознательные психические состояния, то есть психические состояния, в которых сущность осознает [3].

© Пшенокова И.А., 2022

В теории функциональных систем П.К. Анохина [4] сознание понимается как последний этап преобразования всей информации, поступающей из окружающего мира. Сознание содержится в функциональных системах, представляющих собой организации нейронов, распределенных по разным отделам мозга, но одновременно активизирующихся для выполнения определенных функций.

А.Р. Лурия определил сознание как способность оценивать сенсорную информацию, реагировать на нее критичными размышлениями и действиями и сохранять следы событий в памяти, чтобы прошлые отпечатки или действия могли быть использованы в будущем [5].

К функциональным моделям сознания можно отнести модель глобального рабочего пространства (Global Workspace Theory (GWT) Баарса [6] и ее реализации [7, 8]. Согласно GWT, бессознательные процессы и психические состояния конкурируют за центр внимания, из которого информация «транслируется глобально» по всей системе. Сознание состоит в глобальном вещании и поэтому, по словам Баарса, является важной функциональной и биологической адаптацией. Можно сказать, что сознание создается своего рода глобальным доступом к избранным битам информации в мозге и нервной системе.

Другой подход – модель сознания, сформулированная Стивеном Гроссбергом [9] и основанная на адаптивных резонансах в мозге. Согласно этой модели, конкретные нейронные цепи и системы в разных частях мозга генерируют адаптивные резонансы, которые поддерживают сознательное осознание и знания о внешних сенсорных входах, таких как конкретные виды и звуки, или внутренние сенсорные входы, такие как конкретные эмоции. Каждый резонанс поддерживает фокус внимания на перцептивном или аффективном представлении, которое становится сознательным, и синхронизируется с соответствующими внимательными резонансами, которые позволяют сознательно распознать или знать перцептивное или аффективное событие. Причем сознательные перцептивные, когнитивные и аффективные представления могут резонировать вместе, чтобы люди могли осознавать то, что воспринимают, и субъективный опыт при данном подходе – не что иное, как некоторое конкретное состояние в динамической эволюции нейронных сетей.

Модели сознания следует отличать от так называемых нейронных коррелятов сознания, предложенных Ф. Криком и К. Кохом [10]. Хотя идентификация корреляций между аспектами мозговой активности и аспектами сознания может ограничивать спецификацию нейробиологически правдоподобных моделей, такие корреляции сами по себе не обеспечивают объяснительные связи между нейронной активностью и сознанием. Модели также следует отличать от теорий, которые не предлагают какую-либо функциональную реализацию (например, теории «мышления высшего порядка» Розенталя [11]). Модели сознания ценны именно в той мере, в какой они предлагают такие объяснительные связи [12]. Чтобы какая-то функция в мозге была сознательной, нужно, чтобы выполнение определенной когнитивной задачи сопровождалось внутренним субъективным опытом. Именно наличие приватного опыта является ключевым компонентом, позволяющим сказать, есть сознание или нет. Это более узкое понятие называется феноменальным сознанием (phenomenal consciousness). Самосознание возникает в ходе развития сознания личности, по мере того как она становится самостоятельно действующим субъектом. Самосознание есть не столько рефлексия своего «Я», сколько осознание своего способа жизни, своих отношений с миром и людьми.

Таким образом, не существует общепринятого определения сознания. Однако следует указать, что основные определения можно разделить на два понятия: сознания как опыта и сознания как функции. С точки зрения опыта субъект осознает, когда чувствует зрительные переживания, телесные ощущения, мысленные образы, эмоции [13]. С точки зрения функции сознательный субъект способен обрабатывать доступную информацию [14], интегрировать ее [15] и интроспективно осознавать себя [16], генерировать внутреннюю

речь [17], внутреннюю модель себя и внешней среды [18], а также предвидеть перцептивную и поведенческую деятельность [19] и взаимодействовать через афферентные сенсомоторные связи с внешним миром [20].

В этой статье обобщены модели, которые включают вычислительные, информационные или нейродинамические элементы и предлагают объяснительные связи между нейронными свойствами и феноменальными свойствами.

Некоторые подходы к моделированию искусственного сознания

Без понимания основополагающих аспектов биологического сознания невозможно моделирование искусственного сознания. В книге [21] содержится обзор состояния исследований в области искусственного сознания. В более поздней работе [22] авторы предполагают, что слово «сознание» объединяет два различных типа вычислений обработки информации в мозге: отбор информации для глобального вещания (C1) и самоконтроль этих вычислений (C2). Они утверждают, что современные машины попрежнему в основном реализуют вычисления, которые отражают бессознательную обработку (C0) в человеческом мозге.

Все существующие когнитивные архитектуры полностью удовлетворяют цели, для которой они были созданы, однако ни одна из них не показывает реальное создание личных предпочтений, приобретенных и обработанных через тело и эмоции агента, которые считаются основой для создания потенциального искусственного сознания.

Рассмотрим некоторые подходы к моделированию искусственного сознания.

Теория разума, внимание и роль эмоций – все это важные аспекты в изучении механизмов, лежащих в основе сознания у людей и роботов. В этом контексте в работе [23] предлагается теория, основанная на схеме внимания, в качестве отправной точки для создания сознательного робота – attention schema theory (AST). Согласно AST мозг конструирует не только модель физического тела, но и модель своих собственных внутренних процессов обработки информации. Он строит «схему внимания». Эта схема внимания не только способствует контролю внимания, но и информация, содержащаяся в ней, также имеет последствия для тех утверждений, которые машина может делать о себе. Согласно тому, как мозг может в широком спектре информационных областей проводить постоянно меняющуюся усиленную обработку одних сигналов по сравнению с другими, схема внимания – это набор информации, который описывает акт фокусировки ресурсов на чем-то. Она описывает, что такое внимание, что оно делает, каковы его основные устойчивые свойства, каковы его динамика и последствия, и отслеживает его постоянно меняющееся состояние.

Авторы делают вывод о том, что функционально моделирование себя совпадает с моделированием других на основе опыта, проведенного в [24], согласно которому в правой височно-теменной доле мозга происходит осознание себя, своего «Я» и в этой же доле происходит социальное сознание, т.е. функция приписывания осознания другим людям. Согласно Грациано, на основе AST можно создать робота с богатой внутренней моделью сознания, которая приписывает сознание себе и людям, с которыми взаимодействует, и использует эти знания для предсказания человеческого поведения. В предложенной схеме внимания отсутствует информация о нейронах, синапсах, электрохимических сигналах, нейронной конкуренции и так далее. Авторы данного подхода считают, что конкретные сети в мозге не имеют большого значения, а их метод закладывает концептуальную основу для построения сознания. Поскольку это теория, в которой машина конструирует определенный набор информации и использует его определенным образом.

Методологическая стратегия изучения сознания робота введена в [25] посредством концепции вычислительного коррелята сознания, соответствующего концепции нейронного коррелята сознания в мозге [8]. В работе описывается когнитивная роботизированная система, которая учится выполнять задачи, имитируя движения, показанные

человеком. Робот использует причинно-следственные рассуждения, чтобы сделать вывод о целях человека при выполнении задачи, а не просто имитировать наблюдаемые действия. Когнитивные компоненты интеллектуальной системы робота основаны на нисходящем контроле рабочей памяти, которая сохраняет все причинно-следственные связи, возникающие во время обучения.

По мнению Манзотти и Челлы [26], типичные подходы к сознанию робота, как, например, глобальное рабочее пространство, информационная интеграция, инсценировка, когнитивные механизмы, воплощение, представляют собой старое доброе искусственное сознание (Good Old-Fashioned Artificial Consciousness) и имеют общую концепцию, рассматривающую сознание как эпифеноменальное, т.е. не имеющее физической основы. Авторы делают вывод, что для проектирования искусственного сознания оно не может быть чем-то отличным от физического мира. Поэтому формулируют три допущения: сознание принимается таким же, как и все другие физические свойства вокруг, что-то, что можно измерить, наблюдать; сознание причинно активно и находится в пространстве-времени; оно состоит из материи или энергии. Согласно выдвинутой гипотезе сознание – это сеть объектов и событий, которые благодаря телу с сенсорно-моторно-когнитивными способностями взаимодействуют друг с другом. Сознание – это не внутреннее свойство, а совокупность предметов, которые благодаря телу причинно ответственны за то, что делает тело. Таким образом, авторы смещают фокус изучения сознания робота с внутренних процессов и структур на анализ онтогенетических и эпигенетических отношений, которые организм развивает и поддерживает с внешним миром в течение своей жизни.

В работах [27, 28] подробно приводится структура активного вывода (active inference framework (AIF). По мнению авторов, активная структура вывода является мостом между вычислительной нейронаукой, робототехникой, психологией. Вводится понятие «нейронно реализованная генеративная модель», суть которой состоит в том, что даже если некоторый объект не был известен, но есть какое-либо более раннее знакомство с подобными ему объектами, то можно создать генеративную модель этого объекта, дополняя ее по мере знакомства с настоящим объектом. Генеративная модель дополняется и развивается посредством процесса обучения, который преобразует нейронные сети. Полный объем воплощенной (и опционально дополненной нейронами) генеративной модели в AIF включает в себя восприятие через построение текущего состояния посредством экстеро-, проприо- и интероцепции. На основе полученного восприятия агент выводит действия как переходы из настоящего в предпочтительное будущее состояние, которые минимизируют (вариационную) свободную энергию, ожидаемую при актуализации предпочтительного будущего состояния, что соответствует понятию мотивации вознаграждения. Предполагается, что роботизированная интероцепция будет идентифицировать наиболее значимые величины, такие как потребности в энергии («голод») и телесные повреждения («боль»).

Уинфилд [29] предлагает искусственную теорию разума, которая предоставит роботам новые возможности, связанные с социальным интеллектом, для взаимодействия человека и робота. Автор предполагает, что внутренняя модель, основанная на имитационном моделировании, может предложить новую основу для искусственной теории разума. Внутренние модели снабжают робота моделью самого себя и окружающей среды, включая других агентов. Это делается для того, чтобы робот мог проверить свои возможные действия и предвидеть последствия для себя и других агентов. Эксперименты, представленные в работе, показывают, что робот способен предсказывать последствия своих действий как для себя, так и для одного или нескольких роботов, выступающих в качестве прокси-людей, и выбирать действия на основе либо безопасности, либо этических соображений. Авторы делают вывод, что внутренние модели, основанные на ими-

тационном моделировании, представляют собой вычислительную модель теории моделирования разума и показывают, что такая вычислительная модель предоставляет мощную и реализуемую основу для искусственной теории разума.

Коминелли [30] представил когнитивную систему SEAI (Social Emotional Artificial Intelligence), предназначенную для социальных и эмоциональных роботов, разработанную как био-вдохновленная система с моделью эмоций и способностей к рассуждению. В частности, SEAI представляет собой симуляцию теории сознания Дамасио [31]. Теория сознания Антонио Дамасио состоит из трех этапов: эмоций, проходящих через чувства, чтобы прийти к тому, что он называет «чувствами чувств». Эмоция (или внутреннее эмоциональное состояние) описывается как (бессознательная) нейронная реакция на определенный стимул и реализуется сложным ансамблем нейронных активаций в мозге. Поскольку нейронные активации часто являются подготовкой к (телесным) действиям как следствие внутреннего эмоционального состояния, тело будет изменено во внешне наблюдаемое эмоциональное состояние. Чувство описывается как (все еще бессознательное) ощущение этого состояния тела. Наконец, ядро сознания, или чувство – это то, что возникает, когда организм обнаруживает, что его представление о собственном телесном состоянии (прото-Я) было изменено появлением стимула: он (сознательно) осознает чувство. Боссе [32] разработал модель, основанную на представлениях Дамасио, для моделирования динамики основных механизмов, происходящих в разуме и теле агента. Когнитивная система SEAI была спроектирована, опираясь на теорию Дамасио и модель Боссе, и протестирована в качестве когнитивной системы гуманоидного робота FACE (Facial Automaton for Conveying Emotions). Тестирование выявило некоторые недостатки: контроль гомеостаза отсутствует, физиологические параметры агента являются символическим представлением, такие возможности, как превентивность при принятии решений, еще не рассматривались. Однако SEAI выделяется среди других систем благодаря гибридной концепции. Модульная конструкция архитектуры потенциально позволяет расширение и переносимость системы на любого другого социального робота, просто адаптируя или добавляя низкоуровневые услуги к сенсорному аппарату и двигательной системе конкретного агента. Это может быть сделано, сохраняя «личность», воспоминания, убеждения, опыт и поведенческие черты агента, которые зависят от когнитивной части системы и, следовательно, могут быть перенесены или изменены независимо. Более того, экспертная система, основанная на правилах, которая является ядром когнитивного блока, не накладывает конкретных ограничений на возможности рассуждения и вывода, которыми может быть наделен искусственный агент, что зависит от количества и сложности правил. В проведенных экспериментах SEAI наделила социального гуманоида искусственными эмоциями и чувствами, на которые повлиял контекст, агенту удалось эксплуатировать их для построения мнений о социальном мире, в который он погружен, и на их основе проявлять более сложные социальные навыки.

В работе [33] вводится понятие «самосознание» в рамках разработанной авторами интеллектуальной системы NARS (non-axiomatic reasoning system). В NARS под интеллектом понимается способность системы адаптироваться к окружающей среде и работать с недостаточными знаниями и ресурсами. Для этого система должна быть способна распознавать случайные события, принимать ответы в режиме реального времени, работать с ограниченными ресурсами и учиться на своем опыте в различных областях. Система NARS хорошо согласуется с процессами человеческого разума и, в частности, с теорией Пиаже о том, что ребенок узнает о себе и окружающей среде, координируя чувства (такие как зрение и слух) с действиями (такими как хватание, сосание и шаг), и

постепенно переходит от рефлексивного, инстинктивного действия при рождении к символическим умственным операциям.

Кинучи и Макин [34] предлагают когнитивную нейронную архитектуру для сознательного робота, где основная роль сознания заключается в адаптации на системном уровне. Предлагаемая архитектура, названная «базовая система», основана на двухуровневом дизайне: первый уровень связан с осознанием, привычным поведением и проблемой привязки. Второй уровень связан с общим целенаправленным поведением робота. Базовая система автономно адаптируется к окружающей среде с функциями принятия решения о действии, основанными на оптимизации прибыли системы в каждый момент времени. Основные функции базовой системы состоят из примитивных операций: обнаружение и распознавание объектов из окружающей среды, принятие решения о действиях в пользу распознанных объектов и подготовка следующего действия, включая обучение на системном уровне. Оптимальный план действий рассчитывается за короткое время с использованием рекуррентной нейронной сети на основе Brain-State-in-a-Box (BSB), предложенной Андерсоном [35] и Голденом [36]. Кроме того, предложенная схема обеспечивает функцию мощного обнаружения совпадений паттернов, которая обнаруживает совпадающий паттерн из тысяч параллельных сигналов, представляющих атрибуты объектов. Эта функция предоставляется на основе результатов, связанных с пирамидальным нейроном [37]. Предложенная архитектура, однако, еще не была подкреплена работающей моделью.

В работе [38] в основе когнитивной и вычислительной архитектуры функционального сознания лежит теория глобального рабочего пространства [5] при условии, что она не противоречит ключевым понятиям о природе представлений в мозгу. Это условие касается того, что представления в мозгу находятся «на месте» («in situ»). Это означает, что они действуют (по крайней мере частично) всегда как одно и то же представление в каждом экземпляре когнитивных процессов, в которых они участвуют.

В предложенном подходе глобальное рабочее пространство представляет собой нейронную доску, в архитектуре которой происходят обработка и создание концептуальных структур (например, отношений между словами в предложении или между визуальными особенностями объектов).

Представления «in situ» не входят в глобальное рабочее пространство, а подключаются к нему путем активации созданных концептуальных структур. Несколько представлений могут конкурировать, в результате чего одна и соответствующая ей концептуальная структура временно будут доминировать в рабочем пространстве. И именно доминирующее представление «in situ», выбранное в рабочем пространстве, становится основой для функциональной формы сознания путем (непрерывного) «процесса явных или неявных запросов и ответов». По мнению авторов, такой процесс соответствует когнитивной обработке и сознанию доступа, как в человеческом мозге.

В [39] авторы вводят модель памяти, основанную на нейрофизиологических данных, которая учитывает многие аспекты, такие как постоянство объекта и эпизодическая память. Модель памяти состоит из трех основных частей: сети идентификации (WHAT), сети локализации (WHERE) и сети префронтальной рабочей памяти (WORKING MEMORY). Согласно результатам, авторы утверждают, что предложенная модель памяти может работать как когнитивная карта, поддерживающая элементарные операции планирования. Важным аспектом модели является то, что механизмы, которые заполняют ощущения для генерации восприятия, могут быть отделены от сенсорного ввода и работать изолированно. Предполагается, что робот вместе с механизмами для более развитой системы сенсорной обработки и выбора действий будет иметь необходимое когнитивное оборудование для создания базовой формы сознания – по крайней мере в той

степени, в которой она может быть проверена в поведенческих экспериментах. Фундаментальным аспектом этой модели является то, что сознание не является чем-то, что должно быть добавлено в когнитивную систему. Это происходит естественным образом, как только система памяти будет способна заполнять сенсорную информацию и производить переходы памяти. Это создаст внутренний мир, который может быть использован как для интерпретации внешнего вклада, так и для поддержки мыслей, оторванных от текущей ситуации.

В работе [40] обосновывается концепция самосознания робота. Основываясь на представлениях об окружающей среде, которые объединяют восприятие и действие, внедряются два основных элемента, которые, по мнению авторов, являются необходимым фактором самосознания: самооценка – это способность «знать, что я знаю» и сознательно использовать это знание в выборе действия и мета-рассуждения – это обсуждение собственных рассуждений. В статье предложена глобальная когнитивная архитектура, которая состоит из модуля сенсорного восприятия и моторики (содержит врожденный набор перцептивных способностей к восприятию окружающей среды (зрительное восприятие и проприоцепция), сенсомоторного учебного модуля (обрабатывает доступные входные данные, чтобы обнаружить и узнать, какие действия доступны роботу в текущей среде) и модулей пространственного рассуждения, знаний и базы знаний (генерируют и хранят символьные данные о воспринимаемой среде). В работе приведено тестирование нескольких частей этой архитектуры, используя различные наборы модулей. Однако еще предстоит провести валидацию глобальной архитектуры.

Поскольку сознание является богатым биологическим феноменом, вполне вероятно, что достаточно точная научная теория сознания потребует специальной детализации функциональных моделей. Модели сознания, рассмотренные в этой статье, различаются с точки зрения их уровня абстракции, а также по аспектам феноменального опыта, которые они предлагают объяснить. Однако в настоящее время ни одна модель сознания не представляется достаточной для полного учета многомерных свойств сознательного опыта. Более того, хотя некоторые из этих моделей приобрели известность, ни одна из них еще не была принята в качестве окончательной или даже в качестве основы для построения окончательной модели.

Заключение

Выше представлены только некоторые подходы к моделированию искусственного сознания. О преимуществах одних методов перед другими говорить еще рано. Модели сознания различаются с точки зрения их уровня абстракции, а также по аспектам феноменального опыта, которые они предлагают объяснить. Однако в настоящее время ни одна модель сознания не представляется достаточной для полного учета многомерных свойств сознательного опыта. Эта область ИИ активно развивается, и еще нет единой теории основополагающих принципов и методов создания интеллектуальных систем, обладающих сознанием, способных к пониманию своих действий и целей, а также самосознанию.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Petukhov V.V. *Lektsii po obshchey psikhologii* [Lectures on General Psychology]. Moscow: Izdatelstvo Moskovskogo universiteta [Moscow University Publishing House], 1997. 597 p. (In Russian)

2. Carruthers P. Higher-order theories of consciousness, in The Stanford Encyclopedia of Philosophy, ed. Zalta E. N. Metaphysics Research Lab; Stanford University, CA. 2016.

3. Van Gulick R. Consciousness: in The Stanford Encyclopedia of Philosophy, ed. E.N. Zalta (Metaphysics Research Lab, Stanford University). 2018.

4. Anokhin P.K. *Uzlovyye voprosy teorii funktsional'nykh sistem* [Key questions of the theory of functional systems]. Moscow: Nauka, 1980. 203 p. (In Russian)

5. Luria A.R. *Yazyk i soznaniye* [Language and consciousness]. Ed. E.D. Khomskaya. Moscow: Izdatelstvo Moskovskogo universiteta [Moscow University Publishing House], 1979. 320 p. (In Russian)

6. Baars B.J. In the Theather of Consciousness. The Workspace of the Mind. Oxford: Oxford University Press. 1997. 88 p.

7. Shanahan M.P. A cognitive architecture that combines internal simulation with a global workspace. Conscious. Cognit. 15. 2006. Pp. 433–449. DOI: 10.1016/j.concog.2005.11.005.

8. Franklin S., Madl T., D'Mello S. et al. LIDA: a systems-level architecture for cognition, emotion, and learning. IEEE Trans. Auton. Ment. 2014. Vol. 6. No. 1. Pp. 19–41. DOI: 10.1109/TAMD.2013.2277589.

9. Grossberg S. Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. Neural Networks. 2017. Vol. 87. Pp. 38–95. https://doi.org/10.1016/j.neunet.2016.11.003

10. Crick F., Koch C. The problem of consciousness. Scientific American. Special edition. 2002. Vol. 12. No. 1. Pp. 11–17.

11. Rosenthal D. Consciousness and Mind. Oxford, Clarendon. 2005. 400 p.

12. Seth A.K. Explanatory correlates of consciousness: Theoretical and computational challenges. Cognitive Computation. No. 1. 2009. Pp. 50–63.

13. Chalmers D. Facing up to the problem of consciousness. Journal of Consciousness Studies. 1995. Vol. 2. No. 3. Pp. 200–219.

14. Dehaene S., Lau H. and Kouider S. What is consciousness, and could machines have it? Science. Vol. 358. 2017. Pp. 486–492. DOI: 10.1126/science.aan8871.

15. Tononi G. Consciousness as integrated information: a provisional manifesto. Biol. Bull. Vol. 215. 2008. Pp. 216–242. DOI: 10.2307/25470707.

16. Floridi L. Consciousness, agents and the knowledge game. Mind Mach. No. 15. 2005. Pp. 415–444. DOI: 10.1007/s11023-005-9005-z.

17. Morin A. Possible links between self-awareness and inner speech. Journal of Consciousness Studies. Vol. 12. No. 4-5. 2005. Pp. 115–134.

18. Holland O. A strongly embodied approach to machine consciousness. Journal of Consciousness Studies. Vol. 14. No. 7. 2007. Pp. 97–110.

19. Hesslow G. Conscious thought as simulation of behaviour and perception. Trends Cogn. Sci. Vol. 1. No. 6. 2002. Pp. 242–247. DOI: 10.1016/S1364-6613(02)01913-7.

20. O'Regan J.K., Noë A. A sensorimotor account of vision visual consciousness. Behav. Brain Sci. Vol. 24. No. 5. 2001. Pp. 939–973. DOI: 10.1017/S0140525X01000115.

21. Chella A., Manzotti R. Artificial Consciousness. Andrews UK Limited. 2013. 281 p.

22. Esser S., Lustig C., Haider H. What triggers explicit awareness in implicit sequence learning? Implications from theories of consciousness. Psychological Research. 2021. https://doi.org/10.1007/s00426-021-01594-3

23. Graziano MSA. The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness image. Front. Robot. AI. Vol. 4. No. 60. 2017. https://doi.org/10.3389/frobt.2017.00060

24. Graziano M.S., Guterstam A., Bio B.J. et al. Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. Cogn. Neuropsychol. Vol. 37. No. 3-4. 2020. Pp. 155–172. DOI: 10.1080/02643294.2019.1670630.

25. Reggia J.A., Katz G.E., Davis G.P. Humanoid Cognitive Robots That Learn by Imitating: Implications for Consciousness Studies. Consciousness in Humanoid Robots. 2018. DOI: 10.3389/frobt.2018.00001.

26. Manzotti R., Chella A. Conscious Machines: A Possibility? If So, How? Journal of Artificial Intelligence and Consciousness. 2020. Vol. 07. No. 02. Pp. 183–198. https://doi.org/10.1142/S2705078520710022

27. Linson A., Clark A., Ramamoorthy S. et al. The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. Consciousness in Humanoid Robots. 2018. https://doi.org/10.3389/frobt.2018.00021

28. Chang A., Biehl M., Yu Y. et al. Information closure theory of consciousness. Frontiers in Psychology. 2020. Vol. 11. https://doi.org/10.3389/fpsyg.2020.01504

29. Winfield A. «Why Did You Just Do That?». Explainability and Artificial Theory of Mind for Social Robots. Frontiers in Artificial Intelligence and Applications. Vol. 335. 2020. Pp. 8–10. DOI:10.3233/FAIA200892.

30. Cominelli L., Mazzei D., De Rossi D.E. SEAI: Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind. Front. Robot. AI. 2018. DOI: 10.3389/frobt.2018.00006.

31. Damasio A. Feeling & Knowing: Making Minds Conscious. Pantheon. 2021. 256 p.

32. Bosse T., Heyselaar E.S. Linking Theory of Mind in human-agent interactions to validated evaluations: Can explicit questionnaires measure implicit behaviour? : In Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents. 2021. Pp. 120–127. New York, NY: Association for Computing Machinery. DOI: 10.1145/3472306.3478343.

33. Pei Wang, Patrick Hammer, Hongzheng Wang. An Architecture for Real-Time Reasoning and Learning. AGI 2020: Artificial General Intelligence. 2020. Pp. 347–356. DOI: 10.1007/978-3-030-52152-3_37.

34. Kinouchi Y., Mackin K.J. A Basic Architecture of an Autonomous Adaptive System With Conscious-Like Function for a Humanoid Robot. Front. Robot. AI. 2018. DOI: 10.3389/frobt.2018.00030.

35. Anderson J. The Ersatz Brain Project: A brain-like computer architecture for cognition. IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing. 2012. DOI: 10.1109/ICCI-CC.2012.6311125.

36. Golden M. Statistical Machine Learning. CRC Press. 2020. 506 p.

37. Stuart G., Spruston N. Dendritic integration 60 years of progress. *Nat.* Neurosci. 2015. 18. Pp. 1713–1721. DOI:10.1038/nn.4157.

38. Frank van der Velde. In Situ Representations and Access Consciousness in Neural Blackboard or Workspace Architectures. Front. Robot. AI. 2018. DOI: 10.3389/frobt.2018.00032.

39. Balkenius C., Trond A. Tjøstheim, B. Johansson et al. The missing link between memory and reinforcement learning. Frontiers in Psychology. 2020. Vol. 11. Pp. 34-46. https://doi.org/10.3389/fpsyg.2020.560080

40. Raja Chatila et al. Toward Self-Aware Robots. Front. Robot. AI. 2018. DOI: 10.3389/frobt.2018.00088

Информация об авторе

Пшенокова Инна Ауесовна, канд. физ.-мат. наук, зав. лаб. «Интеллектуальные среды обитания», Институт информатики и проблем регионального управления – филиал Кабардино-Балкарского научного центра РАН;

360000, Россия, Нальчик, ул. И. Арманд, 37-а;

pshenokova_inna@mail.ru, ORCID: https://orcid.org/0000-0003-3394-7682

MSC: 68T42 Original article

BASIC METHODS AND APPROACHES TO ARTIFICIAL CONSCIOUSNESS MODELING

I.A. PSHENOKOVA

Institute of Computer Science and Problems of Regional Management –
branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences
360000, Russia, Nalchik, 37-a I. Armand street

Annotation. The article provides a brief analysis of some methods and approaches to the modeling of artificial consciousness. Some approaches to the definition of consciousness in philosophy, psychology and neurobiology are presented. In particular, functional and neurobiological models of consciousness are considered. Some approaches to artificial consciousness modeling are presented. The considered models fully satisfy the purpose for which they were created, however, none of them yet shows the real creation of personal preferences acquired and processed through the agent's body and emotions, which are considered the basis for the creation of a potential artificial consciousness. This area of general artificial intelligence is actively developing and there is still no unified theory of the fundamental principles and methods for creating intelligent systems that are conscious, capable of understanding their actions and goals, as well as self-awareness.

Keywords: artificial intelligence, consciousness, artificial consciousness, robotics, simulation

The article was submitted 11.02.2022 *Accepted for publication 10.03.2022*

For citation. Pshenokova I.A. Basic methods and approaches to artificial consciousness modeling. *Izvestiya Kabardino-Balkarskogo nauchnogo centra RAN* [News of the Kabardino-Balkarian Scientific Center of RAS]. 2022. No. 2 (106). Pp. 72–81. DOI: 10.35330/1991-6639-2022-2-106-72-81

Information about the author

Pshenokova Inna Auesovna, Candidate of Physics and Mathematics sciences, Head of Laboratory «Intellectual Habitats», Institute of Computer Science and Regional Management Problems – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 37-a I. Armand street;

pshenokova_inna@mail.ru, ORCID: https://orcid.org/0000-0003-3394-7682