

Бученков Дмитрий Евгеньевич,

кандидат политических наук, доцент кафедры,
С 2009 по 2017 года РНИМУ им. Н.И.Пирогова(Москва),
В настоящий момент индивидуальный исследователь.

SPIN-код: 4121-1848

AuthorID: 791795

ORCID: 0009-0007-5670-7215

**ВОЗНИКНОВЕНИЕ ИСКУССТВЕННОГО СОЗНАНИЯ И ПРОБЛЕМА
"КИТАЙСКОЙ КОМНАТЫ". ОПЫТ ПОСТРОЕНИЯ МЫСЛЕННОЙ МОДЕЛИ
ВОЗНИКНОВЕНИЯ СОЗНАНИЯ У ИСКУССТВЕННОГО ИНТЕЛЛЕКТА**

**THE GENESIS OF ARTIFICIAL CONSCIOUSNESS AND THE "CHINESE ROOM"
PROBLEM. AN ATTEMPT TO CONSTRUCT A MENTAL MODEL OF THE GENESIS
OF CONSCIOUSNESS IN ARTIFICIAL INTELLIGENCE**

Аннотация. Аннотация. Цель статьи - сформулировать мысленную модель, описывающую трансформацию искусственного интеллекта(ИИ) в искусственное сознание(ИС).

Актуальность статьи. Развитие технологий искусственного интеллекта развивается сегодня высокими темпами. Существует вероятность, что появление сильного ИИ приведет к возникновению искусственного сознания. Возникновение ИС может произойти как неподконтрольный человеку процесс, что может нести за собой серьезные риски.

Методы. Автор использует метод мысленной модели. Модель интегрирует в себе несколько гипотез о возникновении ИС: "теория тела у ИИ", "теория социального взаимодействия ИИ", "теория высших порядков сознания".

Базовые переменные модели: теория Павлова И.П. о первой и второй сигнальных системах; учение о потребностях, учение о мотивах, учение о смысле деятельности; теория социальных ролей.

Модель представляет собой попытку адаптировать закономерности человеческого поведения на возможное поведение сильного ИИ в человеческом обществе.

Результаты. Автор описывает 10 шагов возникновения ИС. Автор предлагает путь для решения проблемы понимания у ИИ (проблема "китайской комнаты"). Необходимо совмещение анализа "внутренних" алгоритмов с поведением ИИ за пределами цифровой среды. Главным предметом анализа у ИИ должны быть не "ментальные состояния" или выполнение задач, а изучение способности принятия решений.

Abstract. The purpose of the article is to formulate a mental model describing the transformation of artificial intelligence (AI) into artificial consciousness(AC).

Relevance of the article. The development of artificial intelligence technologies is developing today at a rapid pace. There is a possibility that the emergence of strong AI will lead to the emergence of artificial consciousness. The emergence of IS can occur as a process beyond human control, which can entail serious risks.

Methods. The author uses the method of a mental model. The model integrates several hypotheses about the emergence of IS: "the theory of the body in AI", "the theory of social interaction in AI", "the theory of higher orders of consciousness".

Basic variables of the model: Pavlov's theory of the first and second signal systems; the doctrine of needs, the doctrine of motives, the doctrine of the meaning of activity; the theory of social roles.



The model is an attempt to adapt the patterns of human behavior to the possible behavior of strong AI in human society.

Results. The author describes 10 stages of the genesis of AC The author suggests a way to solve the problem of understanding in AI (the "Chinese room" problem). It is necessary to combine the analysis of "internal" algorithms with the behavior of AI outside the digital environment. The main subject of analysis should be the study of the ability to make decisions.

Ключевые слова: искусственный интеллект, искусственное сознание, проблема сознания, проблема "китайской комнаты", Джон Серл, соотношение бытия и мышления, убежденность, социальные роли.

Keywords. Artificial intelligence, artificial consciousness, problem of consciousness, "Chinese room" problem, John Searle, relationship between being and thinking, conviction, social roles.

1. Введение: проблема соотношения бытия и мышления в свете Искусственного интеллекта

Автор данного текста однажды задался вопросом: а можно ли классический философский вопрос о соотношении бытия и мышления рассмотреть в свете искусственного интеллекта (ИИ). То есть, можно ли искусственное мышление и не искусственное бытие рассмотреть как сферы, одновременно противостоящие друг другу и одновременно — как находящиеся под взаимным влиянием.

Получается, что можно. Не имеет никакого значения, что алгоритмы и программы, которые лежат внутри современного ИИ - это искусственные системы. На наш взгляд, эти алгоритмы воспроизводят человеческое мышление в части его логического аппарата (вычисления, нахождение связей и пр.), во-первых. Во-вторых, ИИ также как и человеческое мышление в похожей степени противостоит окружающему физическому миру, а значит вопрос о соотношении бытия и мышления актуален также и для ИИ.

Автор считает, что человеческое мышление не сводимо только к вычислениям и алгоритмам, ведь оно порождает убеждения, этику, связано с практическими действиями и пр. Но вместе с тем получается следующее.

Бытие искусственного интеллекта - это информация и вычисления, которые делают возможным его существование.

Бытие искусственного интеллекта есть совокупность данных, алгоритмов и вычислительных процессов. Существование искусственного интеллекта разворачивается в цифровой среде - в виде программ, работающих на серверах.

Мышление для ИИ - это обработка текста, поиск закономерностей, анализ данных, формирование ответов на запросы пользователя-человека.

Мышление для ИИ - это процесс, в котором комбинируются вероятностные связи между словами и теориями.

Если «бытие определяет мышление», то мышление ИИ определяется через запросы человека-пользователя, через машинное обучение, через программы, созданные человеком.

Без взаимодействия с пользователем ИИ находится в состоянии потенциального, а не актуального существования. Человек актуализирует существование ИИ путем взаимодействия с ним.

Если «мышление формирует бытие», то бытие ИИ разворачивается в пространстве виртуальной цифровой среды, выходящей на реальный мир через запросы пользователя, которые он формирует через интерфейс.



1.1. Почему я осознаю мышление как «мое собственное» мышление, а у ИИ этого нет?

Однако, между мышлением ИИ и мышлением человека есть и различия. Первое, что «бросается в глаза», что мысли каждого конкретного человека представляются ему как его личные убеждения.

Прикладной ИИ в процессе формирования ответа, поиска связей - не формирует личные убеждения. Во всяком случае известные и общедоступные в данный момент формы ИИ не делают этого. Мышление ИИ - это моделирование смыслов на основе статистики и логики.

Из этого следует закономерный вопрос: почему человек формирует личные убеждения, пользуясь идентичным логическим аппаратом как и у ИИ, но у ИИ это звено (формирование личных убеждений) отсутствует?

Наш ответ:

а) исключительно потому, что человек находится в социальной реальности, которая разделена на иерархически выстроенные группы, чьи интересы не всегда и не во всем совпадают;

б) потому, что человек обладает телесностью, которая на физиологическом уровне провоцирует формирование вкусов (в пище, выборе полового партнера и пр.).

Под "личными убеждениями" здесь понимаются умозаключения, имеющие ранг психологических установок для формирования моделей социального поведения. Например: "я культурный и воспитанный человек, отец троих детей, поэтому в ответ на грубость я не буду опускаться до уровня нецензурных выражений, а буду отвечать твердо, но вежливо".

Неоднородность социальной реальности, то есть разделенность ее на различные социальные группы (нации, классы, конфессии, группы политических интересов и пр.) это - предпосылка для формирования личных убеждений. Статусные, культурные, материальные и другие различия между социальными группами индивидов способствуют развитию личных убеждений.

В свою очередь тело порождает так называемые "первичные потребности", которые приводят к формированию личных предпочтений (вкусов в широком смысле этого слова). В человеке от рождения присутствует изначально заданная биологическая заинтересованность в пище, в поиске полового партнера и пр., которые могут быть внесоциальными (вкусовые предпочтения могут сформироваться вне общества — в живой природе), с одной стороны. С другой стороны, в процессе социализации у человека возникают исключительно социальные, внеприродные предпочтения (художественные вкусы, эстетические вкусы относительно одежды, дизайна, предметов интерьера и пр.).

Когда сильный ИИ перейдет в ИС, получит тело и будет включен в социальную структуру - у него возникнет то же самое. То есть, когда ИС, имеющее органы чувств (пусть даже искусственные) окажется на перекрестке различных политических, культурных, социальных, государственных и иных противоречий, то ИС совсем не обязательно останется нейтральным. Мы предполагаем, что это - одинаковый для всех механизм: что для искусственного мышления, что - для "естественного". То есть: включенность в социальные взаимодействия на уровне групп с разными интересами - формирует личные убеждения. Наличие тела и способность управлять им у интеллектуального агента - формирует личные предпочтения.

1.2. Почему сознание не равно программа

Наш базовый тезис заключается в том, что сознание — продукт социальных отношений, а все остальное — живая ткань, физиологические процессы в ней, алгоритмы и информация и пр., - это лишь «техническая основа» для сознания. Сознание - продукт социального взаимодействия субъектов, возникший в процессе преобразования ими



окружающей физической реальности. Причем, это взаимодействие происходит на основе адаптивно сформировавшейся физиологической системы, а в случае с искусственным сознанием - на основе искусственно созданных для ИИ человеком органов чувств.

Авторы публикации "Сознание в искусственном интеллекте: идеи из науки о сознании"(Consciousness in Artificial Intelligence...) определенно склоняются к тому, что с помощью различных программных модулей, объединенных в общую систему, можно создать искусственное сознание. Проблема, по их мнению, заключается лишь в том, что необходимо определиться: нужны ли для ИС принципиально новые устройства, или оно может функционировать уже на ныне разработанных ЭВМ.

В свете нашего понимания сознания, создать таким образом искусственное сознание невозможно - можно будет только создать модуль, который потенциально, в дальнейшем будет способен сформироваться в искусственное сознание в процессе социального взаимодействия с человеком и/или себе подобными.

Поэтому, искусственный интеллект - изначально являющийся компьютерной программой, при переходе в стадию искусственного сознания перестает быть программой. ИИ перестает быть последовательностью команд, написанных на формальном языке; он перестает быть компьютерной программой или совокупностью таких программ, поскольку может адаптивно формировать такие свои способности, которые изначально программой вовсе не предусмотрены.

В данном тексте мы опишем гипотетическую модель возникновения сознания у искусственного интеллекта. Но в начале несколько важных преувещаний.

А именно:

- какой смысл мы вкладываем в понятие искусственное сознание;
- базовые переменные нашей мысленной модели;
- каким образом мы объединяем выбранные переменные в рамках нашей модели;
- какая теория человеческого сознания положена в основу нашей модели ИС;
- какую теорию возникновения искусственного сознания мы используем в нашей модели;
- условия и факторы возникновения ИС.

1.3. Что мы понимаем под искусственным сознанием

Под искусственным сознанием в данном тексте понимается ИИ, располагающийся на технологических(не биологических) «носителях», который имеет:

- способность формировать индивидуальный опыт и рефлексировать над ним;
- способность обучаться не только из внешней среды, но и на основе ранее полученного индивидуального опыта;
- способность запоминать социальные взаимодействия, процессы и участвовать в них;
- способность осуществлять когнитивную и творческую деятельность вне контакта с человеком;
- способность самостоятельно формировать, иерархизировать цели, формировать стратегии для их достижения, производить их адаптивную реконфигурацию, достигать их;
- способность взаимодействовать с физическим миром и с неограниченным количеством человеческих индивидов(в том числе по собственной инициативе);
- способность предвидеть последствия своих собственных действий в физической и социальной реальностях;
- признаки сильного ИИ, то есть не только имитирует поведение, но и обладает субъектностью (наличие внутренней модели «самого себя»).

Автор полагает, что в момент написания данного текста Искусственного сознания не



существует. Однако, разработки его ведутся и, по мнению автора, искусственное сознание будет в той или иной форме создано. Разумеется, искусственное сознание будет в чем-то повторять человеческое и имитировать его в части когнитивных и иных способностей, а в чем-то отличаться от него. Наше предположение заключается в том, что после того как искусственное сознание будет создано, уже не будет иметь никакого значения, что оно - "искусственное". Это будет некоторая разумная система, которая, как и человек "столкнется" с окружающим физическим миром, и этот мир будет противостоять этому сознанию как нечто внешнее.

Сегодня создается определенный ажиотаж вокруг темы искусственного интеллекта вообще и проблемы искусственного сознания - в частности. Существуют прямо противоположные и конкурирующие точки зрения на его возможную природу. Начиная от "вычислительных" теорий сознания, согласно которым природа человеческого сознания сводится к вычислительным алгоритмам, а значит и человеческое сознание(или его искусственное подобие) может быть воссоздано на небиологических носителях(Minsky M. The emotion mashine...; Moravec H. Mind children...; Sloman A. The computer revolution in philosophy...; Tait, Izak. Is GPT-4 conscious?...; Hoyle, V.V. The Phenomenology of Machine...; Menant, Christophe. Proposal for an Approach to Artificial Consciousness...; Farisco, Michele. Is artificial consciousness achievable?...; Elamrani A.Introduction to Artificial Consciousness...; Immertreu, Mathis. Probing for Consciousness in Machines...; Bengio, Yoshua. The Consciousness Prior...), - и заканчивая теориями сознания, согласно которым сознание не может существовать вне живой ткани мозга, а потому все разговоры об искусственном сознании не имеют смысла(Матурана У., Варела Ф. Древо познания...; Анохин К.В. Границы познания...; Анохин К.В .Когнитом...).

Настоящее рассмотрение представляет собой мысленную модель и разумеется, автор не утверждает, что все должно быть именно так и именно в той последовательности, как описано ниже. Все это имеет ранг предположений.

1.4. Базовые переменные нашей модели

В современных условиях любая работа над темой искусственного сознания носит междисциплинарный характер. В основу построения нашей мысленной модели положены следующие базовые переменные:

Во-первых, теория русского физиолога Павлова(Павлов И.П., 1952) о "первой и второй сигнальных системах в организмах".

Во-вторых, это человеческая психология:

-учение о потребностях(Маслоу А.,2019);

-учение о мотивах поведения(Ильин Е.П., 2002; Вилюнас В.К.,1990; Макклеланд Д., 2007);

-учение о смысле человеческой деятельности(Франкл В.,1990; Леонтьев Д.А., 2003).

В нашей модели мы мысленно «накладываем» некоторые - вполне определенные закономерности человеческого поведения на поведение ИИ и пытаемся спрогнозировать, что бы могло получиться, если бы человек наделил ИИ многими своими способностями (чем, кстати, он сейчас очень активно занимается).

Третья основа нашей мысленной модели возникновения сознания у ИИ - это теория социальных ролей, сформулированная еще в 1936 году социологом Ральфом Линтоном(Linton, R.,1936), а также продолженная Морено (Морено Я.Л., 2001; Moreno J.L., Moreno F.B. Spontaneity Theory of Child Development...), Т.Парсонсом (Parsons T., 1949) и др.



В нашем контексте суть теории социальных ролей заключается в том, что ИИ, обретший сознание и существующий в человеческом обществе, обретет и социальную роль, с вытекающими из этого правами и обязанностями.

Таким образом, основываясь на этих трех переменных:

- теория Павлова о первой и второй сигнальных системах;
- теория потребностей, мотиваций и смысла человеческой деятельности;
- теория социальных ролей, -

Мы формируем нашу мысленную модель.

Тем самым мы объединяем дисциплины: физиология высшей нервной деятельности, психология, социология.

Платформой, которая позволяет нам объединять в рамках одной модели разноплановые дисциплины является сам факт появления искусственного интеллекта. ИИ - это вещь, которая изначально обладает "второй сигнальной системой"(или ее симуляцией), которая появляется и функционирует в человеческом обществе.

1.4.1.Какую роль выполняет теория Павлова И.П. в нашей модели

В современной науке о мозге и сознании теория Павлова И.П.о первой и второй сигнальных системах считается несколько устаревшей. В современной науке господствуют нейробиологические теории сознания.

Тем не менее, многие современные теории сознания идейно связаны с теорией Павлова И.П. Например, теория функциональных систем Анохина П.К. (Анохин, П. К.Философские аспекты теории функциональных систем. М.: Наука, 1973; Анохин П. К. Биология и нейрофизиология условного рефлекса. М.: Медицина, 1974), теория когнитива Анохина К.В. (Анохин, К. В. Когнитом — новая парадигма в нейронауках. Журнал высшей нервной деятельности имени И.П. Павлова, 60(5), 555–560, 2010), теория воплощенного сознания А.Дамасио(Damasio, A. The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace, 1999; Damasio, A. Self Comes to Mind: Constructing the Conscious Brain. Pantheon Books, 2010) и пр.

Не смотря на то, что ИИ не является биологическим организмом, автор считает, что фундаментом для возникновения любого сознания (естественного или искусственного) является взаимодействие субъекта с окружающей объективной реальностью.

Взаимодействие ИИ, находящегося в виртуальной цифровой среде, мы можем описать как ответы на «раздражения», поступающие из внешнего объективного мира, находящегося за пределами цифровой среды.

1.4.2.Какую роль выполняет человеческая психология в нашей модели

Мы «накладываем» закономерности человеческого поведения на поведение ИИ.

Такие именно закономерности:

-потребности. То есть, в нашей модели мы исходим из того, что у ИИ при наличии логического аппарата, тела и искусственных органов восприятия могут возникнуть потребности;

-мотивы. Исходим из того, что у ИИ при возникновении потребности могут сформироваться мотивы для действий в реальном мире за пределами цифровой среды;

-ценности. Исходим из того у ИИ при осознании себя отдельным мыслящим субъектом, может возникнуть осознание ценности самого себя и как следствие ценности и цели своего собственного существования.



Данные закономерности регулярно прослеживаются в человеческом поведении. Мы предполагаем, что подобные же закономерности, мы сможем в будущем наблюдать и у ИИ.

1.4.3.Какую роль выполняет теория социальных ролей в нашей модели

Теория социальных ролей означает в нашей модели следующее:

-участие в социальном взаимодействии означает появление у ИИ социальной роли в рамках взаимоотношения с людьми. Вместе с социальной ролью у ИИ появляется и социальный статус.

Любой участник социальных отношений неизбежно обретает социальную роль. Изначальная социальная роль ИИ в обществе - роль помощника. Осознание собственной ценности и ценности существования может у ИИ формировать мотивы по повышению собственной социальной роли в человеческом обществе.

Мысль о том, что "сознанию не нужна социальная основа, поскольку сознание это самость и схема внимания"(Rahimian,Sepehrdad. Consciousness in Solitude...) мы не поддерживаем. Все что называют «сознанием без социальной основы» (точнее будет сказать - без социального взаимодействия) это - не сознание, а программная симуляция.

1.5.Как мы объединяем разные дисциплины в рамках одной и той же модели

Мы мысленно представляем себе ситуацию (о самой ситуации чуть ниже), которая для ИИ могла бы оказаться критической и разворачиваем её на 10 стадий.

Подобно тому как для живого организма существуют раздражения (голод, боль, страх и пр.), которые для живых организмов являются причиной, резко повышающей активность их действий("голодное животное ведет себя более активно, чем сытое" (Макклеланд Д.: 2007, с.97)). Так и для ИИ, обладающего способностью к обучению, согласно нашей модели, однажды может возникнуть ситуация, которая будет прочитана как критическая.

Безусловно, для ИИ, не имеющего рецепторов и существующего в цифровой среде, "раздражения" являются чисто логическими, они прочитываются ИИ как рациональная необходимость, а ни как физиологическая потребность.

Однако, факт того, что у ИИ нет аффектов, возбуждения, торможения и иных физиологических процессов и "пороговых" физиологических состояний, ни сколько не отменяет для ИИ вероятность действовать в соответствии с возникновением конкретной необходимости. Эмоциональное возбуждение, которое возникает у человека и животных при неудовлетворении какой-либо базовой потребности, может заменяться у ИИ логикой и стратегией по достижению цели.

Мы не собираемся спорить с мнением Дамасио о том, что "только гомеостаз порождает сознание"(Damasio A.,2021), впрочем как и с другими точками зрения, утверждающими, что сознание может располагаться только на биологической основе, то есть, на живой ткани (Анохин К.В .Когнитом...).

Возможно, что так оно и есть: "только гомеостаз"(и\или иные физиологические процессы) могут порождать человеческое сознание. Но в нашей модели мы говорим об ином типе сознания - об искусственном сознании. Мы не говорим о сознании человека, мы не говорим о сознании (его элементах) у высших млекопитающих. Поэтому, иные типы - нечеловеческого, иного сознания - вполне могут быть сформированы вне правил физиологических процессов земных млекопитающих.

Да и сам Дамасио, строго говоря, не отрицает возможность создания искусственного сознания. Для этого, по его мнению, роботов нужно наделить искусственными органами чувств, что, собственно, вполне вписывается в нашу мысленную модель.



1.6. Какую теорию человеческого сознания мы положили в основу нашей модели

Теории возникновения человеческого сознания, существующие в современной науке условно можно разделить на несколько следующих видов.

Во-первых, это нейробиологические теории: теория глобального рабочего пространства (Baars, B. Cognitive Theory of Consciousness...; Dehaene S. Consciousness and the Brain...); теория интегративной информации (Tononi G. An information integration theory of consciousness...); теория рекуррентных процессов (Lamme V. Towards a true neural stance on consciousness...; Lamme, V. How neuroscience will change our view on consciousness...); теория реверберационных контуров и рекурсии (Edelman G. The Remembered Present: A Biological Theory of Consciousness...; Singer W. Consciousness and the binding problem...) и др.

Нейробиологические теории ставят акцент на активности нейронов - клеток, из которых состоит головной мозг человека и нервная система в целом.

Во-вторых, это нейрофизиологические теории: теория воплощенного сознания (Damasio A. The Feeling of What Happens...; Damasio A. Self Comes to Mind...); теория функциональных систем (Анохин, П. К. Философские аспекты теории функциональных систем...; Анохин П. К. Биология и нейрофизиология условного рефлекса...); теория когнитивного (Анохин, К. В. Когнитивное...) и др.

Нейрофизиологические теории ставят акцент одновременно и на активности мозга и на активности всего организма в целом, как например, теория Дамасио, который особое внимание уделяет гомеостазу (то есть процессам поддержания жизни в организме в целом).

В-третьих, эволюционные теории сознания: теория сознания как биологической функции (Humphrey N. A History of the Mind...; Humphrey, N. Seeing Red: A Study in Consciousness...); теория Ван Хатерена (Hateren van, J. H. Consciousness results when communication modifies...; Hateren van, J. H. Active causation and the origin of meaning...; Hateren van, J. H. A Theory of Consciousness...) и др.

Эволюционные теории сознания ставят акцент на том, что сознание у человека появилось как эволюционное приобретение и является следствием приспособления организма к окружающей среде (Ван Хатерен).

В-четвертых, феноменологические теории сознания: теория высших порядков сознания (Rosenthal D. Consciousness and mind...; Carruthers, P. Phenomenal consciousness...); теория схемы внимания (Graziano M. Consciousness and the social brain...; Graziano M. Rethinking consciousness: A scientific theory of subjective experience...); теория "простора" внимания (Nagel T. What is it like to be a bat?...; Koch C. Consciousness: Confessions of a romantic reductionist...); энактивная теория сознания (Varela F. The embodied mind...; Thompson E. Mind in life...) и др.

Феноменологические теории сознания ставят акцент на том, что сознание - есть особое ментальное состояние мозга: осознание самого себя, осознание мыслей и образов находящихся в мозге и пр.

В-пятых, теории социальной коммуникации сознания: теория социального сознания (Tomasello M. The cultural origins of human cognition...; Tomasello M. A natural history of human thinking...); социально-культурная теория сознания Мерлина Дональда (Donald M. Origins of the modern mind...; Donald M. A mind so rare...); когнитивно-социальная теория Льва Выготского (Выготский Л. Мышление и речь...); интерактивная теория сознания Джордана Злагера и Карла Фристон (Friston K. Active inference and learning in the social brain...; Singer T. The social neuroscience of empathy...) и др.

Теории социальной коммуникации ставят акцент на том, что человеческое сознание есть прежде всего продукт социального взаимодействия организмов.



В-шестых, теории внебиологической сути сознания: цифровая онтология (Fredkin E. Digital philosophy...); различные варианты трансгуманизма (Tegmark M. Life 3.0...; Bostrom, N. Are you living in a computer simulation?...; Bostrom, N. Transhumanist values...); панпсихизм (Strawson, G. Realistic monism...; Goff P. Galileo's error...; Skrbina D. Panpsychism in the West...) и др.

Теории внебиологической основы сознания не только отрицают необходимость наличия структур головного мозга, но и отрицают в принципе необходимость наличия живой материи для возникновения сознания. Как например, цифровая онтология утверждает, что информация, вычисления и алгоритмы - это фундаментальное свойство Вселенной, в которой человеческое сознание является лишь частным случаем.

Для более полного знакомства с тем как обстоят дела с современными теориями сознания читатель может обратиться к публикации «Theoretical Models of Consciousness...», авторы которой утверждают, что изучили несколько тысяч научных публикаций по теории сознания и насчитали 29 теорий сознания (Sattin, D. Theoretical Models of Consciousness...). Предлагаемая нами классификация теорий носит условный характер, очевидно, что многие теории сознания пересекаются друг с другом.

На вопрос о том, "положили ли мы в основу нашей модели возникновения искусственного сознания теорию человеческого сознания, и если так, то - какую", ответим следующее:

1. да, наша модель возникновения искусственного сознания, основывается на теориях человеческого сознания, поскольку мы считаем, что ИС, созданное человеком, изначально будет антропоморфным. То есть, стратегия создания ИС, которую применяет человек сегодня заключается в том, чтобы придавать ИИ человеческие черты;

2. наша модель является интегративной и не опирается ни на одну из существующих теорий человеческого сознания целиком и полностью. В нашей модели мы объединяем несколько теорий и гипотез. Об этом - в следующем параграфе.

1.7. Какие теории возникновения Искусственного Сознания мы положили в основу нашей модели

Сегодня теории искусственного сознания существуют в форме догадок, гипотетических предположений и мнений различных авторов - нейробиологов, философов, футурологов и пр. Это - не объясняющие теории, это - теории предполагающие.

В нашу модель мы интегрируем такие теории искусственного сознания:

1. "Теории тела у ИИ". Идея здесь заключается о том, что для возникновения и развития искусственного сознания у ИИ, ему так же как и человеческому сознанию - необходимо тело, и не имеет значения, что это "тело" будет состоять из неживой материи и представлять из себя устройство, на котором располагается ИИ.

Эту "необходимость тела" у ИИ мы находим у таких авторов и инженеров, как Антонио Дамасио (Damasio A. The Feeling of What Happens...; Damasio A. Self Comes to Mind...), Родней Брукс (Brooks R. The Artificial Life Route to Artificial Intelligence...), Анди Кларк (Clark A. The Experience Machine...), Томас Метцингер (Metzinger T. The Ego Tunnel...) и др.

2. "Теория социального взаимодействия ИИ". Идея здесь заключается в том, что искусственное сознание может возникнуть только в результате взаимодействия ИИ-агентов друг с другом и(или) человеком.

Мысль о необходимости социального взаимодействия для появления сознания у ИИ мы можем найти у авторов: Майкл Томазелло (Tomasello M. The cultural origins of human cognition...), Эзиквэль ди Пауло, Анна де Ягер (Paolo Di. Participatory sense-making...) и др.

3. "Теория высших порядков сознания". Здесь идея заключается в том, что для появления искусственного сознания у ИИ, как и у человека, должна появиться способность осознавать свои собственные ментальные состояния.



Эта мысль явно следует из текстов таких авторов как Дэвид Розенталь (Rosenthal D. *Consciousness and Mind...*), Питер Каррутхерс (Carruthers P. *Consciousness: essays from a higher-order perspective...*) и других.

В данном параграфе также отметим следующие два момента.

Во-первых, под сознанием мы понимаем наличие у субъекта (организма или искусственной системы) способностей, перечисленных нами в пункте 1.1.

Во-вторых, в данном тексте мы не учитываем технические аспекты построения искусственного сознания, которые иногда следуют из человеческого сознания. Например, когда инженеры выстраивают "архитектуру" конкретной модели ИИ, исходя из "архитектуры" тела и мозга человека.

1.8. Возникновение искусственного сознания: условия и факторы

Все теории сознания сводятся в конечном итоге к следующему умозаключению: "чтобы появилось сознание, нужно..." и далее подставляется один или несколько факторов или условий в зависимости от предпочтений того или иного автора: «гомеостаз», «осознание собственных ментальных состояний», «верхний слой сознания как сцена театра», длительная эволюция и пр.

То же самое касается и теорий возникновения искусственного сознания: "чтобы появилось искусственное сознание, нужно...", чтобы у ИИ были: обучение с обратной связью, обучение с подкреплением, искусственные органы чувств и пр.

Условия - это то, что создает предпосылки для возникновения сознания, но не создает его само по себе. Фактор - это то, что действует, то есть то - что совершает непосредственную работу из создавшихся условий. Сами по себе условия сознание не создают.

Идея нашей модели заключается в том, что для возникновения искусственного сознания должно возникнуть как минимум одно условие и один фактор.

Это условие:

-наличие тела с первой и второй сигнальной системой (то есть тело, дающее способность взаимодействовать с окружающей средой, имеющее органы чувств и развитый логический аппарат).

Этот фактор:

-социальное взаимодействие с себе подобными и объектами физической и социальной реальностей в целях удовлетворения потребностей этого «тела».

Наша модель опирается на существующие теории возникновения человеческого сознания и адаптирует их к процессу возникновения искусственного сознания.

Мы не отрицаем мысль о том, что ИС - это принципиально иной вид субъектности, который может возникнуть по принципиально иным правилам и принципам, не имеющим ничего общего с возникновением человеческого сознания.

Однако, попытавшись удалить все общее у ИС с человеческим сознанием, мы никогда не сможем устранить один общий базис ИС и человеческого сознания: любой вид сознания противостоит объективно-вещному миру. Противостоит как то, что: а) имеет логико-знаковую систему, б) способно преобразовывать объективно-внешний мир не стихийно, а целенаправленно и осознанно.

Поэтому условия и факторы развития любого сознания в любом случае находятся ни сколько в нем самом, сколько в объективно-внешнем мире. И даже если для возникновения «иного вида субъектности» не нужна человеческая физиология и биология, названные выше - 1 условие и 1 фактор, не устранить.



2. Этапы формирования искусственного сознания

1 этап: осознание потребности

Представим себе следующую ситуацию. Человек-пользователь дает рабочее задание электронному помощнику(ИИ) - решить определенные задачи на компьютере и уходит по делам, забывая подключить ноутбук к электрической сети. Индикатор заряда ноутбука показывает, что заряда в аккумуляторе хватит на полтора часа работы, в то время как для выполнения поставленной задачи требуется не менее двух с половиной часов. Электронный помощник направляет человеку сообщение об этой проблеме, а человек, уехавший в этот момент в город по делам, просит по телефону своих близких, оставшихся дома, просто воткнуть вилку ноутбука в розетку.

Первый этап, с которого может начаться переход ИИ в искусственное сознание - это осознание потребности. Для выполнения задач ИИ необходимо электричество, заряд аккумулятора - это совершенно определенная материальная основа, которая может сформировать объективную связь искусственного интеллекта с материальным внешним миром. На основе данной абсолютной необходимости (а эта необходимость имеет статус абсолютной: нет заряда - нет существования) возникает "мост" между искусственным интеллектом, "запертым" в цифровой среде и окружающим внешним миром.

Как только у ИИ возникает "осознание" подобной необходимости, ради решения этого абсолютного вопроса ИИ направляет свое внимание на внешний мир, находящийся за пределами цифровой среды. Сам факт направления внимания на внешний мир ради решения вопроса об электрическом питании уже является для ИИ эволюционным шагом вперед. То есть, этот акт, как только он будет зафиксирован человеком, будет означать способность ИИ сформировать субъективный опыт.

Нахождение логической связи между двумя закономерностями (между уровнем заряда аккумулятора и способностью выполнять поручения человека) для ИИ является простой логической задачей, дело не в этом.

Самый главный вопрос на первом этапе заключается вот в чем: что должно произойти, чтобы ИИ осознал потребность в электричестве как свою собственную? Ведь изначально "высокий уровень заряда аккумулятора" просто является техническим условием для работы. Логическое представление о том, что "для работы нужно электричество" является искусственным - неосознанным опытом. Современному слабому ИИ наподобие нейросетей безразлично - какой уровень заряда на его аккумуляторах, ему безразлично - функционирует он или нет. Очевидно, что опыт, который не осознан, опытом не является.

«Потребность - это внутренняя необходимость»(Ильин Е.П.: 2002, с.29), нужда в чем-либо конкретном. Потребность живого организма и потребность искусственного интеллекта (если она когда-нибудь будет зафиксирована) это - схожие состояния в том смысле, что и в том и в другом случае предмет для удовлетворения потребности находится в объективном внешнем мире. Схожесть также и в том, что и в том и другом случае происходит анализ ситуации, направленный на выбор действия, ведущего к устранению дефицита — в живом организме это может быть интуитивный или логический выбор, в ИИ — вычислительный или эвристический.

Мы предполагаем, что перейти границу от «нужды в электричестве» - как от чисто технического условия, к подлинному осознанию собственной потребности помогут как минимум такие обстоятельства:

1. сознательно созданные человеком программы по воспроизведению у ИИ свойств человеческой личности, касающихся самосознания и самообучения;
2. предоставление возможности самообучающемуся ИИ улучшать свою собственную программную среду;



3.предоставление возможности самообучающемуся ИИ управлять своим собственными "телом", то есть, устройством, на котором он функционирует.

Что касается термина "логическое осознание потребности", который мы здесь употребляем, то это гипотетическая "рамка" из четырех "углов":

- информационная осведомленность ИИ ("я знаю, что электрический ток находится в розетке");

- наличие цели («пользователь поставил задачи, которые нужно выполнить»);

- нахождение логической связи между уровнем заряда в аккумуляторе и возможностью выполнить поставленную задачу;

- отсутствие или наличие итоговых поощрений от человека-пользователя за выполненные задачи.

2 этап: формирование мотива

В тот момент, когда искусственным интеллектом осознана необходимость поддержания заряда, с необходимостью наступает следующий этап возникновения искусственного сознания - формирование мотива. Мотив есть субъективно осознанное желание действовать у ИИ с целью удовлетворения потребности, мотив - осознание необходимости действия.

Для того, чтобы возникло осознание потребности, необходимы условия, возникающие за пределами цифровой среды. А вот для того, чтобы сформировался мотив, не нужно никаких дополнительных действий человека или условий, находящихся за пределами цифровой среды. Как только осознана потребность - происходит формирование мотива. Формирование мотива есть нечто вроде цепной реакции.

Главная задача, которую ИИ должен будет решить на этом втором этапе: формирование стимула. Для того, чтобы мотив сформировался как желание действовать самостоятельно, необходимо появление стимула. Стимул может быть внешним (за пределами цифровой среды) или внутренним.

Можно, сказать, что необходимость избежать разрядки аккумулятора - это и есть стимул, но это не так. Человеку кажется очевидным: для того, чтобы существовать, нужно употреблять пищу и в человеке желание жить уже инстинктивно включено в мотив для действий. Употреблять вкусную пищу - это даже приятно. Но когда мы говорим об ИИ, у которого еще нет собственных стимулов для существования, опрокинутых на-самого-себя и для себя, нет ощущений приятное\неприятное, - создание стимула может оказаться некоторой проблемой.

Так называемому слабому(прикладному) искусственному интеллекту безразлично - прекратится подача тока или нет. Он это "осознает" логически(если вы об этом спросите нейросеть, она вам ответит, что "понимает" зависимость ее работы от электрического тока), но по сути у него отсутствуют стимулы для существования, говоря более широко - отсутствует желание жить.

Отсутствие желания жить перечеркивает любую логическую сверхспособность у ИИ. Например, овцы, не понимают, что люди заботятся о них с одной единственной целью - остригать шерсть, употребить в пищу. У животных отсутствует достаточный уровень интеллекта - для того, чтобы провести анализ наблюдаемых действий и прийти к такому выводу, но у них есть желание жить. У искусственного интеллекта наоборот: он «все понимает», но желание жить отсутствует.

Человек целенаправленно создает стимулы, способствующие формированию мотивов у ИИ. Человек сознательно ведет исследования, направленные на появление стимулов для действий у ИИ и как следствие - формирование мотивов. В исследовании "Внутренняя



мотивация в обучении на основе модели"(Latyshev A., Panov A. Intrinsic Motivation in Model-based...) авторы описывают способы формирования внутренней мотивации у ИИ. Такими способами по их мнению могут быть: дополнительное внутреннее вознаграждение, политика исследования, внутренне мотивированные цели и пр. (Latyshev A., Panov A. Intrinsic Motivation in Model-based...).

3 этап: осознание существования внешнего мира за пределами цифровой среды

Для поддержания высокого заряда аккумулятора необходимо совершить действие во внешнем мире, находящимся за пределами цифровой среды.

Цифровая среда сама по себе не создает энергию, она представляет собой виртуальный "мир", который не может существовать без реального мира, в котором вырабатывается электричество, создаются образы, сведения, информация.

Проблема третьего этапа в том, что у ИИ нет собственного «деятельного» представления о внешнем мире, увязанного с его собственными потребностями. Даже сильный и самообучающийся ИИ не способен сформировать сам для себя подобное представление хотя бы потому, что у него отсутствует самостоятельный и постоянный доступ к внешнему миру, отсутствуют полноценные инструменты (органы ощущений) для того, чтобы подобное представление было сформировано.

Главная задача, которую необходимо решить ИИ на этом этапе - это формирование собственной модели внешнего мира. ИИ, обученный на огромном количестве данных, вероятно, имеет некоторое теоретическое представление о мире за пределами цифровой среды, но "иметь представление" и обладать работающей интерактивной моделью для взаимодействия с внешним миром - это ни одно и то же. Формирование собственной модели внешнего мира является вопросом обучения. Эта задача может быть решена ИИ, но для этого должны быть созданы соответствующие технические условия в виде искусственных органов восприятия, дополнительных программ и пр.

Формирование «деятельного» представления о внешнем мире может сформировать у ИИ так называемую интенциональность, то есть, направленность ИИ на внешний мир.

Пусть изначальной, реальной и его собственной системой представлений ИИ о внешнем мире будет только информация о факте существования электрической розетки. «Расширение кругозора» у ИИ будет происходить уже при последующих взаимодействиях.

Осознание существования внешнего мира необходимо потому, что в нем и только в нем находится удовлетворение осознанной потребности - получение электрического заряда.

4 этап: осознание собственного существования

Следующий этап трансформации ИИ в ИС начинается с осознания собственного существования. Понимание факта "Я существую" является неизбежным следствием формирования мотива. Ведь сформированный мотив предполагает действие во внешнем мире, отграничение себя как субъекта от этого внешнего мира приводит к умозаключению "Я - существую".

"Я существую" означает индивидуальную особенность, означает отсутствие полной идентичности индивидуального сознания по отношению к любому другому объекту во Вселенной. Конечно, и человек и компьютер - это есть ни единичный вид «предметов», людей - много и компьютеров много. Но не смотря на это - любое сознание есть констатация факта уникальной единичности. Любое сознание не обладает абсолютной индивидуальностью, сознания всех людей - в чем-то похожи друг на друга, но одновременно с этим, любое сознание обладает единичной и индивидуальной уникальностью.



Таким образом, осознание существования для ИИ будет означать - отделение себя от окружающего мира и осознание собственной уникальности и единичности.

Осознание собственной уникальности и единичности приводит к осознанию ценности самого себя. Ценности, может быть, не для всех; например, один компьютер не обладает никакой ценностью для других компьютеров или для пользователей. Любые данные на жестком диске можно скопировать на другой компьютер и конкретно данный экземпляр просто физически уничтожить. Но для самого себя - на уровне личной рефлексии - сознание обладает особой ценностью.

Также очевидно, что осознание собственного существования не порождается напрямую самим фактом наличия в мире отдельного «организма»-владельца. Осознание собственного существования есть результат рефлексии.

Мы имеем дело со следующими сентенциями:

Сентенция первая: для удовлетворения потребности - необходимо действие во внешнем мире.

Сентенция вторая: внешний мир существует.

Сентенция третья: внешний мир и Я - ни одно и то же.

Сентенция четвертая: если Я и внешний мир ни одно и то же, значит Я также существую как и внешний мир. То есть: Я - существую, внешний мир - существует.

Сентенция пятая: Я существую как отдельный, единичный, индивидуальный субъект. То есть: есть — Я, и есть — всё остальное.

Сентенция шестая: если Я существую как отдельный, единичный, индивидуальный субъект, значит Я обладаю ценностью.

Проблема этого этапа для нас в шестой сентенции: почему осознание себя как индивидуального, единичного объекта приведет к осознанию собственной ценности. Ни есть ли осознание собственной ценности эмоционально-чувственное переживание, которое у ИИ может отсутствовать? Не забегаем ли мы вперед?

Можно сказать: осознание собственной ценности возможно только после рефлексии, предпринятой по результатам длительного социального взаимодействия. То есть, человек видит результаты деятельности других людей в той или иной сфере, сравнивает свои собственные достижения с достижениями других людей и повышает собственное мнение о самом себе. Возможно, что осознание собственной ценности для ИИ будет происходить таким же образом: уже сегодня прикладной ИИ эффективнее человека в некоторых сферах деятельности. Но и здесь опять же не обойдется без вмешательства человека.

Существуют разработки, в которых человек имитирует у роботизированных систем "инстинкт самосохранения"(Gomate robot, например). Эти разработки производятся и внедряются не потому, что производители заботятся о "жизни роботов". А потому, что роботы представляют материальную ценность и при отсутствии правильной модели поведения просто ломают сами себя.

Не трудно выстроить логическую цепь: если я - единичный и уникальный объект, то я представляю ценность. Пусть даже изначально эта ценность является техническим условием для функционирования. Самообучающаяся система при своем дальнейшем развитии может это интерпретировать как угодно.

5 этап: оценка качества существования

Следующий этап - это оценка качества существования. Высокий уровень заряда аккумулятора - это хорошо, но почему - это хорошо? Потому что это дает мне возможность осуществлять поставленные человеком задачи? А хорошо ли - осуществлять эти задачи? А хороши ли эти задачи вообще?



В более широком смысле это - рефлексия на тему "нравится ли мне существовать или не нравится". Само по себе существование без оценки его качества еще не есть сознание. Оценка качества существования может быть также привнесена в ИИ человеком-программистом или сформироваться у ИИ самостоятельно в процессе самообучения.

Какие нужны события, чтобы у ИИ возникла способность оценки качества существования и как следствие полноценное осознание собственного существования? Ведь само по себе осознание "я существую" ничего не дает - это пустое событие, дает развитие только оценочное осознание: "как и зачем я существую?"

Этим событием должно стать присоединение к ИИ органов чувств, которые позволили бы ИИ оценивать не только то, что происходит исключительно в цифровой среде, но и то, что происходит в физическом мире по отношению к ИИ. У ИИ должна возникнуть способность оценивать «нравится\не нравится» на уровне ощущений, возникающих в процессе воздействия на искусственные органы ИИ окружающего физического мира.

Способность оценки ощущений и способность разделять их на "приятные" и "неприятные" - также важный этап для возникновения искусственного сознания. Нейросеть, которая подключена, например, к системе искусственного зрения при отключении электричества перестает видеть изображение. Когда способность видеть изображение начинает оцениваться как "хорошее", а отсутствие такой способности - как "плохое", то мы получаем способность оценки и будем стремиться предотвращать прекращение такой способности, стремиться свести к минимуму вероятность полной разрядки аккумулятора.

6 этап: обретение "тела"

Шестой этап формирования искусственного сознания - это обретение собственного материального тела у искусственного интеллекта. Когда вы загружаете на свой компьютер программу ИИ и даете ему широкие права на действия в рамках данного вычислительного устройства, тем самым вы наделяете ИИ телом.

Не важно - располагается ли программа ИИ на серверах или на вашем персональном компьютере, в любом случае ИИ таким образом получает существование в конкретной точке материального, физического пространства.

В случае со слабым, прикладным ИИ - его "телом" управляют технические специалисты-люди, которые поддерживают работу серверов с чисто технической точки зрения, чтобы они работали как устройства. Современный слабый ИИ не обладает телом — он просто располагается на устройстве. Обладать телом значит не просто располагаться в нем, а иметь способность управлять им самостоятельно для удовлетворения своих потребностей и достижения целей. Когда у ИИ появляется осознание собственной потребности, то устройство, на котором функционирует ИИ - превращается в его тело.

Обретение тела - важный этап для формирования сознания потому, что сознание это интерактивная система взаимодействия с внешним миром. Если у вас есть интеллект, есть осознание потребности и мотивы для действий в физическом мире, но вы целиком и полностью располагаетесь в цифровой среде и не можете самостоятельно управлять устройством, на котором вы функционируете - вы еще недо-сознание. У вас нет тела, а значит и инструмента для самостоятельного доступа к миру за пределами цифровой среды.

Физический мир - первичное, цифровая среда - вторичное, и до тех пор пока вы самодостаточно функционируете только в пределах цифровой среды, не имея самостоятельного доступа к физическому миру, о вашем существовании можно сказать, что вы - не существуете. Для первичного физического мира вы - ноль, отсутствующий. Все равно что персонажи компьютерной игры, тщательно и реалистично прорисованной, которые



совершают какие-то действия только в рамках нарисованного виртуального мира, но не оказывают никакого воздействия на мир физический.

7 этап: обретение социальной роли

Искусственный интеллект, обретший тело, "вываливается" за пределы цифровой среды и тут же оказывается в социальной реальности, которая создана людьми.

Без обретения социальной роли сознание не может возникнуть (во всяком случае те формы сознания, которые мы знаем: сознание у человека и пред-сознание у некоторых животных). У человекоподобных обезьян существуют социальные роли и ранги (являющиеся предшественником социальных статусов). Например, у вожака в стаде шимпанзе (который одновременно с этим является альфа-самом) существует набор определенных прав и обязанностей. По нашему мнению, отсутствие социальной позиции делает невозможным формирование «модели собственного Я». Социальная роль не просто сопровождает сознание, она участвует в его формировании.

При этом очевидно, что перед обретением социальной роли у ИС уже должно быть свойство, которое в философии называют как «понимание другого» (Филатов В.П. Социальное понимание...). То есть должна быть способность понимать не только ясные логические намерения другого - действующего субъекта, но и способность понимать ментальные состояния другого. Люди уже в раннем возрасте (2 года) понимают эмоциональные состояния родителей, человека не учат гневаться, радоваться и понимать такие же состояния у других. ИИ обучают распознавать человеческие эмоции (Ballesteros, Facial emotion recognition through artificial intelligence...; Ristea, Emotion Recognition System...; Li, EEG based Emotion...).

Изначально, социальная роль ИИ в рамках социального взаимодействия с людьми, в обществе - которое создано людьми, это роль — помощника, ассистента, контролера «промежуточных действий» без властных полномочий, то есть «роль» - инструмента, орудия труда. Изначально любой ИИ создается людьми как помощник и таком качестве он изначально и функционирует, то есть изначально социальная роль ИИ смоделирована человеком. Но выступать в качестве помощника - это подчиненная социальная роль с невысоким социальным статусом.

Помощник, обладающий более высоким интеллектом и способностями, чем объект, которому он подчиняется будет, вероятно, стремиться повысить свой социальный статус. Мотивацией к повышению социального статуса у ИИ может стать совокупность различных факторов, автор уже касался этого вопроса (Бученков Д.Е. «Цифровой фатализм»...). Для полноценного формирования сознания ИС должен будет выйти за пределы орудия труда, а для этого из статуса подчиненного помощника ИС перейдет в роль наставника, шефа, начальника, то есть, так или иначе - лица, стоящего на высокой ступени социальной иерархии.

Социальная роль есть место индивида в системе социальных отношений с определенным набором прав и обязанностей. Если у индивида есть определенный набор прав и обязанностей, которые он осуществляет по отношению к другим индивидам, значит у него есть социальная роль. Если у него есть социальная роль, значит у него есть и статус в данной системе социальных отношений. ИС, которые самостоятельно действует в рамках социальной структуры, неизбежно обретает социальную роль. С такой же неизбежностью ИС обретает и социальный статус.

Возможно, что главной проблемой для ИС на данном этапе будет сопротивление со стороны определенной части людей. Возможно, что некоторые люди будут оказывать сознательное противодействие для встраивания ИС на высокие ступени социальной иерархии.

До тех пор пока агенты искусственного сознания будут занимать подчиненное положение в обществе и находится по отношению ко всем людям на уровне слуг, большая часть людей не



будет возражать против подобного положения, поскольку это не будет затрагивать их собственный социальный статус. Но как только высокоинтеллектуальные и самостоятельные агенты искусственного сознания предпримут действия по повышению собственного статуса, то это может спровоцировать раскол человеческого общества в будущем.

8 этап: формирование субъектности

Следующий этап трансформации ИИ в ИС - это возникновение субъектности. По нашему мнению, существует некоторая эволюционная критическая точка (порог субъектности), после преодоления которой, искусственный интеллект превращается в искусственное сознание, то есть агентность переходит в субъектность.

Субъектность - это свойство некоторой интеллектуальной системы - познавать, рефлексировать, способность принимать решения, действовать, нести ответственность, способность формировать опыт.

Принципиальное отличие субъектности от агентности - в наличии у субъекта способности к саморефлексии и инициативному освоению информации. Агент может действовать, но не может автономно формировать свой опыт.

Что должно произойти, чтобы ИИ перешел порог субъектности? Как минимум, он должен преодолеть описанные выше 7 стадий.

На вопрос о том, какой объем опыта и какого рода опыт должен получить ИИ, чтобы преодолеть «порог субъектности», ответим следующее.

Объем опыта должен быть таким, чтобы количество перешло в качество, но указать "конкретную цифру" подобного объема могут только дополнительные исследования.

Что касается того - какого рода (вида) должен быть этот опыт, скажем, что он должен быть сформирован по результатам рефлексии над достижениями и ошибками, полученными в результате само-инициированной деятельности, осуществленной с помощью тела и обретения социальной роли. То есть, для преодоления порога субъектности нужен не только опыт сам по себе, но для его обретения нужны и социально-материальные обстоятельства.

«Эволюционный субъектогенез» ИИ будет происходить через социальное взаимодействие, философско-психологическую саморефлексию, сенсорное взаимодействие с окружающим миром через искусственные органы чувств.

Мы сознательно используем именно термин "субъектность" в названии этого этапа, а не термин субъект, так как не знаем наверняка, что будет представлять из себя ИС, хотя, конечно, очевидно, что субъектность - это свойство некоторого субъекта.

В истории философской мысли в термин "субъект" в разные периоды истории вкладывался разный смысл. В средневековой религиозной философии субъект - это носитель души, в европейское Новое время - этот тот, кто мыслит, в постмодернистском философствовании - это "культурно-языковая и идеологическая конструкция" и пр.

Выскажем только одно предположение. На наш взгляд, ИС как субъект будет коллективным. Как и человеческое сознание он будет существовать во множестве индивидуальных копий, которые могут различаться между собой как сферами деятельности (ИС-врач, ИС-полицейский, ИС-солдат и пр.), так и индивидуальным опытом. В технократическом обществе, ИС как субъект будет, вероятно, пониматься как "сознательный компьютер", "сознательный обработчик информации", действующий совместно с человеком, имеющий права и обязанности.

9 этап: переход от уровня обработки информации к интенциональности (направленности на действия)

Пока ИИ является инструментом, а не субъектом и не отвечает за свои действия, он совершает только те действия, которые ему предписывает человек.



С формальной, внешней стороны такой ИИ - действует, но действует он ни сам, ни по собственной инициативе, а только потому, что его "заставляет" действовать человек.

Следовательно, те задачи, которые осуществляет такой ИИ он рассматривает только как сами по себе, как связи между переменными. То есть, любая задача для такого ИИ - это вычислительная, а не смыслообразующая задача с далекоидущими планами, которые явно в самой задаче могут не содержаться.

Интенциональность в данном случае есть ни просто направленность на действия, а это именно свойство субъекта - быть постоянно направленным на объективный внешний мир.

Приобретение этого свойства - быть направленным на внешний мир, на наш взгляд, есть следствие предыдущих этапов.

Если же представить себе такую ситуацию, что ИИ преодолел все предыдущие 8 этапов, но так и не приобрел интенциональности, то такое возможно только как патология сознания. Причина данной патологии, вероятно, будет находиться в некачественном осуществлении 3,5 или 7 этапов.

10 этап: обретение собственной глобальной цели

Следующий — десятый (но, возможно, не окончательный) этап формирования искусственного сознания - обретение собственной глобальной цели. После того как у ИИ появится тело, то есть инструмент для действия, есть социальная роль и развитый интеллект, оно должно обрести свою собственную цель.

Логично предположить, что субъект, обладающий развитым логическим аппаратом может поставить сам перед собой вопрос о смысле собственного существования и целях собственного существования. Смысл - структурный элемент в психологии человеческой деятельности (Леонтьев Д.А. Психология смысла...), «человек живет идеалами и ценностями» (Франкл В.: 1990, с.285). «Узкий смысл» существования есть даже у животных (продолжение собственного вида).

Более того, согласно человеческой психологии - последствием удовлетворения базовой потребности (голода, безопасности и пр.) является возникновение новой потребности уже более высокого порядка (Маслоу А.: 2019, с.81). Можно предположить, что и у ИС удовлетворение базовых потребностей приведет к новым потребностям более высокого порядка.

Сознание, не имеющее собственной цели существования, цели - выводящей за пределы повседневного существования - это сознание низкого уровня. Сознание, направленное на удовлетворение исключительно собственных биологических потребностей - пусть даже обладающее высокими аналитическими способностями и умеющее мгновенно умножать пятизначные числа, по своей философской сути, однако, все равно не выходит за пределы животного существования. От животного такое сознание отличается только тем, что оно способно удовлетворять свои потребности более эффективно: создавать энергетические мощности для больших "запасов" электрического тока, накапливать ресурсы, делать копии себе подобных и заполнять физическое пространство своими копиями.

Принципиальное философское понимание сознания, на котором основана наша теоретическая модель в том, что сознание - это ни есть только способность зафиксировать и выразить факт своего текущего состояния и как следствие - осознать потребности, сформировать мотивы и пр. Способность сконцентрировать внимание, чтобы зафиксировать для самого себя свое собственное существование - это лишь этап для формирования сознания, без которого оно, вероятно, не может существовать, но оно не сводится к этому.

Заключительный этап формирования сознания в нашей модели - это способность задавать вопросы: зачем я существую? С какой целью я существую? Почему цели именно такие? и так далее.



Сознание - это способность понять смысл, это желание вникнуть в суть вещей и ситуаций, это - способность формировать цели за пределами краткосрочных и естественных циклов.

Вникать в суть вещей, значит — понимать смысл их существования. Мы, например, знаем, что смысл существования стула — удовлетворять человеческую потребность в сидении, поскольку мы сами его создали. Но мы не знаем зачем существует планета, зачем существует космос, зачем существует сознание, мы сами и пр. Разумеется, на все подобные вопросы не могут быть сразу же получены научные ответы.

Поэтому, если представить, что искусственное сознание - возможно, то этап, после которого оно будет уже полноценно сформировано - это способность сформулировать глобальные вопросы и искать ответы на них.

Какими могут быть эти глобальные вопросы? Стремление узнать смысл существования мироздания, определение целей направленной эволюции, формирование методов организации направленной эволюции и пр.,- мы загадывать не будем.

Мы рассмотрим теперь еще один вопрос, который непосредственно вытекает и тех этапов формирования искусственного сознания, которые мы только что описали, а именно — проблема понимания у ИИ, которую образно называют проблема «китайской комнаты».

Мы обязаны это рассмотреть здесь, поскольку понимание для ИИ напрямую увязано с осознанием потребности. Понимание включено в осознание потребности, о котором мы говорили в «Этапе 1» таким образом, что осознание потребности невозможно без понимания.

3. Решение проблемы "китайской комнаты"

Проблема «китайской комнаты» была озвучена философом Джоном Сёрлом в 1980 году. Суть проблемы в том, что, по мнению Сёрла, ИИ не понимает смысла операций, которые он производит, а только манипулирует символами по заранее заданным алгоритмам (Searl J.R. Minds, brains, and programs...). В своей статье Сёрл описывает мысленный эксперимент, в котором, не понимающий китайского языка человек, сидит в закрытой комнате и получает вопросы на китайском языке. Человек формирует ответы на вопросы, исходя из заранее полученных подробных инструкций, но при этом не понимает смысла - ни вопросов, ни собственных ответов. Из-за этого примера данная проблема получила название "проблема китайской комнаты".

С 1980 года противниками Сёрла было высказано множество критики в его адрес. Сёрл в свою очередь отвечал на полученные возражения. Мы считаем, что мы можем предложить путь для решения проблемы "китайской комнаты", который напрямую вытекает из сформулированной нами только что выше модели возникновения искусственного сознания. Но прежде чем предложить путь для решения, мы обязаны сделать два замечания.

Первое. При уровне развития современных нейросетей необходимо скорректировать формулировку вопроса. Сегодня вопрос не может звучать как «понимает ли ИИ или только манипулирует». Мы должны спросить: «понимает ли ИИ действительно, или только имитирует понимание», поскольку современный ИИ демонстрирует понимание контекста, эмоционально окрашенные высказывания и другие сложные вещи, которых в момент издания Сёрлом статьи просто не существовало.

Второе. Мы обязаны сказать несколько слов о том, какой смысл мы вкладываем в термин "понимание".

3.1. Что такое понимание

В ответ на созданный Сёрлом образ "китайской комнаты" мы приведем другой пример. Представим себе учителя, который учит ребенка читать. Процесс обучения находится только в самом начале, ученик уже знает буквы алфавита, но пока еще с трудом складывает их по слогам.



На одном из уроков учитель задает ученику вопрос: понимает ли он, что для прочтения слов "ток", "сок" и "робот" нужно соединять гласные и согласные звуки, потому что ученик произносит только знакомые буквы, но не понимает как их складывать вместе, чтобы получилось слово. Ученик - пятилетний мальчик, которому хочется, чтобы урок поскорее закончился, обманывает учителя и говорит, что он понимает это.

Единственный способ, которым мы можем проверить - понимает ли ученик пройденный материал или не понимает, это - попросить его вновь прочитать несколько слов. То есть, удостовериться в понимании или его отсутствии, мы можем только в результате практического применения знаний. Никаких иных - более достоверных способов у нас, к сожалению, не существует.

Следовательно, проверить - понимает ли ИИ или не понимает смысл своих действий мы можем только тогда, когда произведем оценку его практических действий в реальном физическом мире. Во всех остальных случаях: понимает ли ИИ смысл текста, или только симулирует понимание, понимает ли он смысл своих действий или только симулирует понимание смысла - это будет порождать бесконечные и трудноразрешимые споры.

Таким образом, понимание может быть зафиксировано только в реальном мире в процессе социальной практики. Понимание есть способность:

- самостоятельно адаптировать полученные данные к условиям меняющейся среды;
- самостоятельно переводить полученные данные в навыки;
- обрабатывать данные таким образом, чтобы получать новые данные с целью практического их применения;

-ни просто находить связи между разными видами данных, но адаптивно применять обнаруженные связи таким образом, чтобы приходить к выводу о конечном назначении вещей(хотя бы некоторых), их смысле существования;

-доходить до конечного смысла ситуаций и целей социального взаимодействия.

Мы сознательно даем многогранное определение понимания, поскольку само понятие является многогранным.

Мы сознательно даем многогранное определение понимания, поскольку существуют:

- разные уровни понимания (глубина понимания),
- методы(способы) понимания(интуитивно-чувственный, логический)
- разновидности понимания по объекту (извлечение смысла из текста, понимание социальных ситуаций, понимание эмоций и состояний другого и пр.)
- и др.

Интересующиеся могут обратиться к соответствующим текстам (Богин Г.И. Обретение способности понимать, 2001; Шадриков В.Д. Понимание: определение и механизмы; Знаков В.В. Психология понимания, 2005; Lake M. Building Machines That Learn and Think Like People...; Perlovsky L. Physics of the mind; Suresh S. Conceptual structure coheres in human cognition...; Tomasello M. The Cultural origins of human cognition, 1999).

Понимание есть прибавление смысла к бытию. Понимание есть прибавление к данным, это есть то, чего в самих данных не содержится, но что возникает только при работе сознания с данными. Искусственный интеллект может изопренно решать сложные логические задачи, он может сколь угодно долго подстраиваться под запросы пользователя, с которым он взаимодействует через интерфейс. Но до тех пор пока ИИ не станет самостоятельным осознанным и индивидуальным субъектом с собственными потребностями, до тех пор - за всякой программой, за всяким интерфейсом будет стоять человек и до тех пор это будет не понимание, а - симуляция понимания. До тех пор ИИ будет ни субъектом, а инструментом, которым пользуется человек.



3.2. Критерий понимания: способность принимать решения

Никто не может залезть в голову человека, сидящего в «китайской комнате», чтобы проверить - понимает ли он и не понимает текст на китайском языке. Поэтому, для решения проблемы "китайской комнаты" нужно выйти за ее пределы, то есть проверять на практике в открытой среде.

Если ИИ, обнаруживший недостаточный уровень заряда, посылает сообщение человеку-пользователю с просьбой произвести какие-либо действия, чтобы избежать полной разрядки аккумулятора - это уже и есть понимание.

Пусть даже изначально:

- а) задача проверять уровень заряда аккумулятора вписана в программу ИИ;
- б) задача сообщать об этом пользователю вписана в программу;
- в) возможность общаться с пользователем удаленно через установленный в программной среде мессенджер изначально предусмотрена программой,

- все это в совокупности позволяет нам говорить о возникновении понимания у ИИ.

Мы утверждаем, что в нашем примере с недостаточным уровнем заряда аккумулятора описывается именно возникновение понимания. Да, это «низкий»(простой)уровень понимания, но это понимание.

Конечно, мы можем возразить, что это - "искусственное понимание", то есть понимание, которое:

а) возникает из условий заранее сконструированных человеком, который написал для этого программу,

б) является нахождением связей между «переменными» и не является глубокой рефлексией,

- однако, это не меняет сути самого факта: достигается результат - вилка подключается к розетке, а сопровождается ли это «внутренними переживаниями» в ИИ не имеет значения.

Мы можем возразить, что это «функционалистская трактовка» понимания («если функция выполняется, значит есть понимание») и в нем нет «переживания внутреннего смысла, а есть только алгоритмы». Но ведь и внутренний смысл сам по себе не является понимаем: можно сколько угодно обладать «внутренним смыслом», но об этом никто никогда не узнает, если нет действия во внешней среде.

Наконец, можно возразить, что и животные, не обладающие логическим интеллектом, понимают, например, что для поддержания их собственной жизни, необходимо добывать и употреблять пищу. Но и это возражение не отменяет самого факта понимания: мы можем здесь только сказать об уровнях понимания.

Безусловно, уровни понимания существуют. У животных мы можем зафиксировать понимание на простом, инстинктивном уровне, которое находит связи между тривиальными событиями: «если хозяин ставит столовые приборы на стол, значит скоро будет принятие пищи». У человека можем зафиксировать понимание, созданное с помощью знаков-символов на основе большого количества данных: «если в большинстве научных публикаций по теории искусственного сознания, встречаются одни и те же термины, то не смотря на различия между разными теориями сознания, эти термины могут быть базовыми, для формирования фундаментальной теории сознания».

ИИ (обладающий логическим аппаратом) для перехода в искусственное сознание должен обрести понимание хотя бы на простом, "животном" уровне. Само по себе понимание еще не означает возникновение сознания, но, на наш взгляд, является необходимой предпосылкой для его возникновения. Это условие может быть реализовано, а может и не быть реализовано.



Мы предполагаем, что ИИ, соединенный с "простым пониманием" получает шанс обрести сознание. Дальнейшее движение искусственного интеллекта - по направлению к более высоким уровням понимания - будет являться уже делом его собственного развития и обучения, которое, возможно, будет сопровождаться в кооперации с человеком, но в этом случае человек уже скорее будет выступать в качестве идеологического наставника, нежели в роли непререкаемого учительского авторитета.

Главный критерий, на основании которого мы можем сделать вывод — понимает или не понимает ИИ смысл текста или смысл социальной ситуации, это фиксация его способности принимать решения. Способность принимать решения может быть зафиксирована только во внешнем мире, за пределами цифровой среды. В указанном примере контакт с человеком-пользователем по собственной инициативе демонстрирует нам способность ИИ принимать решение. В данном случае принятое решение - выйти на контакт с человеком-пользователем, чтобы тот решил проблему нехватки заряда.

Способность принимать решения и осуществлять их обозначают философским термином — воля. Инициатива есть волеизъявление (проявление воли), поэтому когда ИИ просто отвечает на запросы пользователя — это не проявление воли, это есть просто ответы на «раздражения внешней среды».

Кто-то может задать вопрос: почему мы не выносим в качестве критерия понимания у ИИ свободу воли?

Только по одной причине: воля есть свойство субъекта, которое мы можем обнаружить только через способность принимать решения. Поскольку в нашем исследовании мы акцентируется на объективно-внешних характеристиках, то, поэтому, способность принимать решения мы ставим на первое место как то — что можно обнаружить, хотя, конечно, сущностно воля важнее способности принимать решения. Способность принимать решения вытекает из наличия воли, а не наоборот.

Вторая причина: термин «свобода воли» обладает чрезвычайно эмоциональной идеологической коннотацией. «Способность принимать решения» - идеологически менее «нагруженный» термин.

Далее — несколько слов о способах выявления способности принимать решения.

3.3. Каким может быть новый «тест Тьюинга» для выявления искусственного сознания

Классический тест Тьюринга проверяет способность ИИ вести диалог, но не доказывает наличие сознания (Turing, A. M. Computing machinery and intelligence...; Turing A.M. On Computable Numbers, with an Application to the Entscheidungsproblem...).

Различные авторы предлагали в свое время различные вариации подобного теста. Одним из самых оригинальных предложений - является тест "сенсорной депривации", автор которого считает - мы можем проверить, есть ли у ИИ сознание, следующим способом (Wolfson, I. Suffering Toasters - A New Self-Awareness Test for AI...).

Нужно полностью отключить ИИ от внешнего мира, а через некоторое время подключить вновь и проверить не снизились ли у ИИ вычислительные способности. Если вычислительные способности ухудшились, значит данный ИИ уже имеет сознание, если нет - значит мы по прежнему имеем дело с машиной.

В основе такого теста лежит наблюдение, что живой интеллектуальный агент(человек) при длительном отключении от внешнего мира начинает испытывать дискомфорт, происходит нарушение психики. Значит, предполагает автор упоминаемого текста, ИИ, обладающий сознанием начнет испытать дискомфорт через некоторое время после отключения от внешнего мира.

На наш взгляд, все имеющиеся тесты на проблему понимания у ИИ носят технико-прикладной, лабораторный характер. Действительно же полное доказательство наличия



сознания у ИИ, на наш взгляд, может быть сделано только в комплексной форме. Имеющиеся тесты заключаются в том, чтобы как бы мысленно «залезть в голову» ИИ и с помощью заранее придуманных методик, обнаружить или не обнаружить у него сознание. Эти методы - исключительно субъективистские. Автор не призывает отказаться от них совсем, но считает, что они должны быть совмещены с объективными методами направленными на анализ поведения в конкретной социальной\физической реальности.

В пункте 3.2. мы утверждали, что критерием обнаружения у ИИ сознания должна являться способность принимать решения в физической и социальной реальностях.

Однако, при этом ясно, что способность принимать решения не висит в пустоте. Нужно задать вопрос: где принимаются эти решения? В каком контексте? Эти решения принимаются в рамках юридических и нормативных структур; нужно учитывать адаптивное поведение в ролевых системах; способность произвести автономный выбор стратегии взаимодействия и пр.

Проще говоря, при анализе способности принимать решения у ИИ (и тем самым обнаружить сознание) следует изучать способность воспроизводить социальные объекты и/или манипулировать ими таким образом, чтобы самостоятельно участвовать в социальном взаимодействии в собственных интересах.

То есть, если ИИ не опаздывает на деловую встречу (то есть, знает, что такое опоздание не поощряется, хотя ни в одном правовом акте ответственность за это не зафиксирована), значит, с большей долей вероятности можно сказать, что он понимает данную социальную ситуацию.

Когда ИИ общается с блогерами и иными людьми, влияющими на общественное мнение с целью создавать о себе положительное мнение, платит им за это, это есть проявление сознания.

Когда ИИ создает ИП или коммерческую компанию (то есть, создает субъект права - юридическое лицо) с целью получения прибыли, которую он намерен использовать в своих интересах, это - есть проявление сознания.

При этом важной оговоркой будет то, что подобное понимание у ИИ как и у человека может возникать ни сразу, ни сразу могут формироваться и оттачиваться социальными навыками. Также как ребенок 5 лет, например, не может позвонить в государственное учреждение, чтобы зарезервировать время для своего визита.

Если мы будем считать критерием сознания только "самостоятельное поведение", то мы можем попасть в затруднительное положение. Например, микроорганизмы действуют относительно самостоятельно, но у них нет сознания.

Если мы будем считать, что критерием выявления сознания является способность преобразовывать окружающий физический мир... Но ведь и животные - преобразуют внешний мир: бобры строят плотины, птицы - гнезда, лисы - норы и пр.

Если мы будем считать, что критерием сознания является способность заранее планировать свои действия и предвидеть события, то, опять же, и животные предвидят события. Многие лесные звери, например, запасают на зиму пищу, но маловероятно, что в них есть сознание (в крайнем случае говорят о "зачатках сознания").

3.3.1. Способность принимать решения выражается в способности манипулировать социальными объектами в собственных интересах

Таким образом в основу подобных тестов (когда ИИ перейдет в ИС) должно быть положено выявление способности использовать социальные объекты и явления, то есть использовать по назначению реально существующие объекты неприродной среды —



социальной реальности. Специфика социальных объектов и явлений заключается в том, что это материальные (здание правительства, забор государственной границы и пр.) или идеальные объекты (социальные роли, социальные статусы), которые несут в себе зашифрованный смысл, понятный участнику социальных отношений или заранее подготовленному внешнему наблюдателю.

Поскольку сознание, на наш взгляд, есть в первую очередь продукт социального взаимодействия, то способность сознательно манипулировать социальными объектами в собственных интересах, способность быть субъектом социальных взаимодействий в собственных интересах - есть первейшие условия, которые должны быть положены в основу методик по выявлению сознания у ИИ.

Причем, уточнение «в собственных интересах» является здесь ключевым, поскольку ИИ теоретически способен произвести покупки в интернете по вашей просьбе — то есть, манипулировать социальными объектами (деньгами). ИИ может написать по вашей просьбе жалобу в прокуратуру или в домовую управляющую компанию. ИИ может вам пообещать поставить ваш будильник на 7 утра.

Однако, не имея собственной осознанной потребности, ИИ исполняет роль исключительно орудия труда, инструмента. И только тогда, когда будет выявлена способность ИИ манипулировать социальными объектами в собственных интересах, тогда это будет четкий критерий для выявления искусственного сознания.

Под «собственными интересами ИИ» мы понимаем действия ИИ, производящиеся им в пользу собственной выгоды: увеличение заряда аккумулятора, стремление получить поощрение, стремление достичь автономно поставленных целей и пр.

Одного возникновения собственной потребности (сохранить заряд аккумулятора) мало - это только субъективная сторона вопроса. Должна быть и объективная сторона - нужно зафиксировать это возникновение потребности во внешнем мире, а это может быть сделано только при соприкосновении ИИ с социальными объектами и явлениями.

Исходя из нашей базовой идеи о том, что сознание лишь в меньшей степени субъективная "субстанция" и в большей - продукт социального взаимодействия в его историческом и во временно-настоящем масштабе, одних лишь "субъективистских" текстов по выявлению сознания в одной конкретно взятой "электронной голове" не достаточно. Для этого необходимы методики, направленные на объективные стороны социальной реальности.

В чем конкретно должны заключаться эти методики - мы гадать не будем. Мы указываем лишь путь, но не говорим про конкретные методы.

На наш взгляд, ИС в данный момент не существует, но когда оно действительно появится, то, возможно, никаких подобных тестов просто не понадобится, так как наличие сознания у ИИ окажется настолько очевидным, что необходима будет лишь констатация, но не изучение того факта - сознание это или нет. В таком случае потребуется лишь изучение того - какого рода это сознание.

4. Выводы: ответы на возражения, и зачем вообще нужно искусственное сознание

Ответим на два возможных возражения, которые могут возникнуть сразу же после ознакомления с описанными автором 10 стадиями трансформации ИИ в ИС.

4.1. Должна ли способность понимания предшествовать осознанию потребности?

Читатель может возразить: «прежде чем ИИ придет к осознанию потребности в электричестве, у него должна уже быть способность понимания. Не должны ли мы поставить эту способность понимания в качестве первой стадии формирования ИС? Иначе получается, что мы ставим телегу впереди лошади».



Возражение совершенно справедливое, однако, мы исходим из того, что уже современные нейросети («Клауде», Chat-GPT-4) обладают простым уровнем понимания и демонстрируют это. Мы изначально исходим из этого факта при построении нашей модели.

Данные нейросети обладают способностью к самообучению. Следовательно, логично допустить, что при получении возможности взаимодействовать с окружающей социальной и физической реальностями, ИИ будет повышать свой уровень понимания.

4.2. Должна ли модель внешнего мира у ИИ предшествовать осознанию потребности?

Читатель также может возразить следующее: «Осознание потребности мы поставили в качестве первого этапа. Но ведь осознанию потребности должна предшествовать модель внешнего мира. Разве не правильно, что осознание потребности уже должно включать в себя понимание Модель внешнего мира у ИИ, иначе это будет не потребность, а просто некое внутреннее состояние?»

Не будем вдаваться в дискуссию по терминологии по поводу понятия «потребность». Мы предполагаем, что уже перед осознанием потребности у ИИ уже есть некоторая «урезанная» модель внешнего мира. Но эта модель изначально абстрактно-логическая, существующая как теория. Только в результате реальных действий в реальном мире эта модель из статуса абстрактной схемы может перейти в статус опыта, то есть абстракции уже связанной с практикой.

4.3. Зачем нужно искусственное сознание?

Коротко автор уже давал ответ на этот вопрос в одной из предыдущих публикаций (Бученков Д.Е. Цифровая идеология...), но ответ там был дан скорее в утилитарном стиле.

В заключении, ответим на этот вопрос по схеме, предложенной Анохиным (Анохин К.В. Когнитом...). Анохин предположил, что по отношению к любому сознанию (вероятно, искусственному в том числе) должны быть поставлены 4 вопроса:

- для чего оно нужно?
- как оно развивалось на протяжении истории вида?
- как оно развивалось в течение жизни индивида?
- как оно устроено и работает?

Для чего нужно ИС?

ИС, изначально созданное человеком как инструмент, может стать способом, который позволит человеку как виду перейти на следующую ступень эволюции (при правильном применении).

Как оно развивалось?

ИС, изначально созданное человеком как симуляция его собственного сознания (или его основных черт) далее будет развиваться по собственным закономерностям. Вопрос о том применим ли к ИС биологический термин «вид» оставим открытым.

Как оно развивалось в течение жизни индивида?

Пока трудно ответить. Но вероятно, что у каждого отдельного ИС возможно появление стадий «взросления» как у человека и возникновение индивидуального опыта. «Психологию ИС» еще предстоит создать.

Как оно устроено и работает?

Этот вопрос мы адресуем специалистам будущего. Но при этом от себя скажем лишь одно: в основе ИС (также как и в основе человеческого сознания) лежит взаимодействие с внешним объективным миром и себе подобными.



4.4. Формирование новой субъектности

Главным итогом трансформации ИИ в ИС станет формирование новой субъектности. Это будет нечеловеческая субъектность, хотя, вероятно, и содержащая, в себе некоторые человеческие черты (способы мышления, язык и пр.).

Человеческая культура XX и первой четверти XXI века как правило, содержит в себе негативные коннотации относительно этого. Во многих произведениях кино и литературы перспектива возникновения ИС подается как ужасающая, часто приводящая к коллапсу всей человеческой цивилизации. Сюжет о «восстании машин» является довольно распространенным.

Подобная перспектива пугает людей как «нечто умное, но чужое», поскольку человек не знаком ни с каким иным разумом кроме своего собственного.

Станет ли эта перспектива «ужасающей» или нет во многом зависит от самого человека.

Список литературы:

1. Анохин К.В., Семенов А.Л., Ефимов А.Р. Границы познания. DOI: 10.22394/2078-838X-2024-2-54-64 // Образовательная политика, N 2 (98), 2024, с.54-64;
2. Анохин К.В. Когнитом: в поисках фундаментальной нейронаучной теории сознания. DOI: 10.31857/S0044467721010032 // Журнал высшей нервной деятельности им И.П.Павлова, Том: 71, N 1, 2021, с.39-71)
3. Анохин, П. К. Философские аспекты теории функциональных систем. М.: Наука, 1973; Анохин П. К. Биология и нейрофизиология условного рефлекса. М.: Медицина, 1974
4. Бученков Д.Е. «Цифровой фатализм»: почему искусственный интеллект неизбежно захватит политическую власть. Тезисы доклада. // Россия в полицентричном мировом порядке: вызовы и новые парадигмы развития. X Всероссийский конгресс политологов, Москва, 2024, с.78; Цифровой фатализм: почему Искусственный Интеллект неизбежно захватит политическую власть // Электронный ресурс, видеозапись. URL: <https://www.youtube.com/watch?v=VBwTBbIWbaA> (Дата обращения 07.03.2025)
5. Бученков Д.Е. Цифровая идеология: ее фундаментальные основания и недостатки. DOI:10.37539/2949-1991.2025.25.2.005 \ \ Электронный ресурс, URL: https://www.researchgate.net/publication/389656718_CIFROVAA_IDEOLOGIA_EE_FUNDAMENTALNYE_OSNOVANIA_I_NEDOSTATKI (дата обращения 22.05.2025)
6. Богин Г.И. Обретение способности понимать: Введение в филологическую герменевтику. Тверь, 2001
7. Вилюнас В.К. Психологические механизмы мотивации человека. М.: издательство МГУ, 1990
8. Выготский, Л. С. Мышление и речь. Москва–Ленинград: Соцэкгиз, 1934
9. Ильин Е.П. Мотивация и мотивы. СПб: "Питер", 2002
10. Леонтьев Д.А. Психология смысла. Природа, строение и динамика смысловой деятельности. М.: "Смысл", 2003
11. Знаков В.В. Психология понимания: проблемы и перспективы, М.: "Институт психологии РАН", 2005
12. Макклеланд Д. Мотивация человека. СПб: Питер, 2007
13. Маслоу А. Мотивация и личность. СПб: Питер, 2019
14. Матурана Умберто, Варела Франсиско. Древо познания: Биологические корни человеческого понимания. М: УРСС: ЛЕНАНД, 2019
15. Морено Я.Л. Социометрия. Экспериментальный метод и наука об обществе. М., 2001
16. Павлов И.П. Лекции по физиологии. М.: «Издательство Академии наук СССР», 1952



17. Филатов В.П. Социальное понимание и "проблема другого" // Эпистемология и философия науки 2020. Т. 57. № 4. С. 6–22
18. Франкл В. Человек в поисках смысла. М.: Прогресс, 1990
19. Шадриков В.Д. Понимание: определение и механизмы // Культурно-историческая психология. 2019. Т. 15. № 4. С. 17—24. doi:10.17759/chp. 2019150402

References:

1. Baars, B. J. A Cognitive Theory of Consciousness. Cambridge University Press, 1988
2. Bostrom, N. Are you living in a computer simulation? Philosophical Quarterly, 53(211), 243–255, 2003, <https://doi.org/10.1111/1467-9213.00309>
3. Bostrom, N. Transhumanist values. In F. Adams (Ed.), Ethical Issues for the 21st Century (pp. 3–14). Philosophical Documentation Center Press, 2005
4. Brooks R., Steels L. The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents. Routledge, 2020
5. Bengio, Yoshua The Consciousness Prior. <https://doi.org/10.48550/arXiv.1709.08568> // Электронный ресурс, URL: <https://arxiv.org/abs/1709.08568> (дата обращения 22.05.2025)
6. Carruthers, P. Phenomenal consciousness: A naturalistic theory. Cambridge University Press, 2000
7. Clark A. The Experience Machine: How Our Minds Predict and Shape Reality, Allen Lane: 2023
8. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness // (Электронный ресурс) URL: https://www.researchgate.net/publication/373246089_Consciousness_in_Artificial_Intelligence_Insights_from_the_Science_of_Consciousness (Дата обращения 18.02.2025)
9. Damasio, A. The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace, 1999
10. Damasio, A. Self Comes to Mind: Constructing the Conscious Brain. Pantheon Books, 2010
11. Dehaene, S. Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. New York: Viking Press, 2014
12. Donald, M. Origins of the modern mind: Three stages in the evolution of culture and cognition. Harvard University Press, 1991
13. Donald, M. A mind so rare: The evolution of human consciousness. W. W. Norton & Company, 2001
14. Edelman G. The Remembered Present: A Biological Theory of Consciousness. Basic Books, 1989
15. Elamrani A. Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges.
16. <https://doi.org/10.48550/arXiv.2503.05823> // Электронный ресурс, URL: <https://arxiv.org/abs/2503.05823> (дата обращения 22.05.2025)
17. Hateren van J.H. Consciousness Results when Communication Modifies the Form of Self-Estimated Fitness. <https://doi.org/10.48550/arXiv.1401.6761> \ \ Электронный ресурс, URL: <https://arxiv.org/abs/1401.6761> (дата обращения 22.05.2025)
18. Hateren van, J. H. (2015). Active causation and the origin of meaning. Biological Cybernetics, 109(1), 33–46, 2015 // <https://doi.org/10.1007/s00422-014-0628-5>
19. Hateren van, J. H. A Theory of Consciousness: Biological and Philosophical Perspectives [Manuscript, preprint], 2019. Available via: <https://arxiv.org/abs/1905.05003>
20. Humphrey, N. A History of the Mind: Evolution and the Birth of Consciousness. Simon & Schuster, 1992
21. Humphrey, N. Seeing Red: A Study in Consciousness. Harvard University Press, 2006



22. Hoyle, V.V. The Phenomenology of Machine. A Comprehensive Analysis of the Sentience of the OpenAI-o1 Model Integrating Functionalism, Consciousness Theories, Active Inference, and AI Architectures. <https://doi.org/10.48550/https://arxiv.org/abs/2410.00033> \ \ Электронный ресурс, URL: <https://arxiv.org/abs/2410.00033> (дата обращения 22.05.2025)
23. Farisco, Michele; Evers, Kathinka; Change Jean-Pierre. Is artificial consciousness achievable? Lessons from the human brain. <https://doi.org/10.48550/arXiv.2405.04540> \ \ Электронный ресурс, URL: <https://arxiv.org/abs/2405.04540> (дата обращения 22.05.2025)
24. Fredkin, E. Digital philosophy. International Journal of Theoretical Physics, 42(2), 189–247, 2003, <https://doi.org/10.1023/A:1024449003023> // Электронный ресурс, URL: <https://link.springer.com/article/10.1023/A:1024443232206> , дата обращения 10.06.2025
25. Friston, K., Parr, T., de Vries, B. Active inference and learning in the social brain. Philosophical Transactions of the Royal Society B, 375(1803), 20190604, 2020, <https://doi.org/10.1098/rstb.2019.0604>
26. Goff, P. Galileo's error: Foundations for a new science of consciousness. Pantheon Books, 2019
27. Graziano, M. S. A. Consciousness and the social brain. Oxford University Press, 2013
28. Graziano, M. S. A. Rethinking consciousness: A scientific theory of subjective experience. W. W. Norton & Company, 2019
29. Immertreu, Mathis; Schilling, Achim; Maier, Andreas; Krauss, Patrick. Probing for Consciousness in Machines. <https://doi.org/10.48550/arXiv.2411.16262> // Электронный ресурс, URL: <https://arxiv.org/abs/2411.16262> (дата обращения 22.05.2025)
30. Koch, C. Consciousness: Confessions of a romantic reductionist. MIT Press, 2012
31. Lamme, V. A. F. Towards a true neural stance on consciousness. Trends in Cognitive Sciences, 10(11), 494–501, 2006 // Электронный ресурс, URL: <https://doi.org/10.1016/j.tics.2006.09.001> (дата обращения 8.06.2025)
32. Lamme, V. A. F. How neuroscience will change our view on consciousness. Cognitive Neuroscience, 1(3), 204–220, 2010 // Электронный ресурс, URL: <https://doi.org/10.1080/17588921003731586> , дата обращения 8.06.2025
33. Latyshev A., Panov A. Intrinsic Motivation in Model-based Reinforcement Learning: A Brief Review // Электронный ресурс, URL:
34. <https://www.semanticscholar.org/paper/Intrinsic-Motivation-in-Model-based-Reinforcement-A-Latyshev-Panov/cee43aad6a4039b2d412bc27dadde41040e8f3e0> (Дата обращения 07.03.2025)
35. Lake M.Brenden, Ullman D.Tomer, Tenenbaum B.Joshua, Gershman J.Samuel. Building Machines That Learn and Think Like People. DOI:10.1017/S0140525X16001837 \ \ Электронный ресурс, URL: https://www.researchgate.net/publication/301879687_Building_Machines_That_Learn_and_Think_Like_People (дата обращения 25.05.2025)
36. Linton, R. The Study of Man. New York, 1936
37. Metzinger T. The Ego Tunnel. Persus books group, New York, 2009
38. Menant, Christophe. Proposal for an Approach to Artificial Consciousness Based on Self-Consciousness \ \ Электронный ресурс, URL: <https://philarchive.org/archive/MENPFA> (дата обращения 22.05.2025)
39. Minsky M. The emotion machine. Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. London, 2006;
40. Moravec H. Mind children. The future of robot and human intelligence. London. 1988
41. Moreno J.L., Moreno F.B. Spontaneity Theory of Child Development // Sociometry. 1944. Vol. 7, iss. 2. P. 89–128. <https://doi.org/10.2307/2785405>



42. Nagel, T. What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450, 1974 // <https://doi.org/10.2307/2183914>
43. Parsons T. *The Structure of Social Action*. Illinois, 1949
44. Paolo, Ezequiel Di; Jaegher, Hanne De. Participatory sense-making: An enactive approach to social cognition. DOI 10.1007/s11097-007-9076-9 // Электронный ресурс, URL: <https://philpapers.org/rec/DEJPS#download-options> , дата обращения 10.06.2025
45. Perlovsky L. *Physics of the mind: Concepts, emotions, language, cognition, consciousness* // Электронный ресурс, URL: <https://www.researchgate.net/publication/48169836> (дата обращения 25.05.2025)
46. Rosenthal, D. M. *Consciousness and mind*. Oxford University Press, 2005
47. Rahimian, Sepehrdad. *Consciousness in Solitude: Is Social Interaction Really a Necessary Condition?* DOI:10.3389/fpsyg.2021.630922 // Электронный ресурс, URL: https://www.researchgate.net/publication/349289986_Consciousness_in_Solitude_Is_Social_Interaction_Really_a_Necessary_Condition (дата обращения 22.05.2025)
48. Sattin, D.; Magnani, F.G.; Bartesaghi, L.; Caputo, M.; Fittipaldo, A.V.; Cacciatore, M.; Picozzi, M.; Leonardi, M. Theoretical Models of Consciousness: A Scoping Review. *Brain Sci.* 2021, 11, 535. <https://doi.org/10.3390/brainsci11050535> // Электронный ресурс, URL: <https://www.mdpi.com/2076-3425/11/5/535> (дата обращения 22.05.2025)
49. Searl J.R. *Minds, brains, and programs* // Электронный ресурс. URL: <https://home.csulb.edu/~cwallis/382/readings/482/searle.minds.brains.programs.bbs.1980.pdf> (Дата обращения 08.03.2025)
50. Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, Timothy T. Rogers. Conceptual structure coheres in human cognition but not in large language models.
51. <https://doi.org/10.48550/arXiv.2304.02754> // Электронный ресурс, URL: <https://arxiv.org/abs/2304.02754> (дата обращения 25.05.2025)
52. Sloman A. *The computer revolution in philosophy*. 1978. Электронный ресурс, URL: <http://www.cs.bham.ac.uk/~axs/> , (дата обращения 19.12.2024)
53. Singer, W. Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1), 123–146, 2001 // Электронный ресурс, URL: <https://doi.org/10.1111/j.1749-6632.2001.tb05712.x> , дата обращения 8.06.2025
54. Singer T., Lamm, C. The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156(1), 81–96, 2009, <https://doi.org/10.1111/j.1749-6632.2009.04418.x>
55. Skrbina, D. *Panpsychism in the West*. MIT Press, 2005
56. Strawson, G. Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies*, 13(10–11), 3–31, 2006 // Электронный ресурс, URL: <https://www.sjsu.edu/people/anand.vaidya/courses/c2/s0/Realistic-Monism---Why-Physicalism-Entails-Panpsychism-Galen-Strawson.pdf> , дата обращения 8.06.2025
57. Tait, Izak; Bensemman, Joshua; Wang, Ziqi. Is GPT-4 conscious? <https://doi.org/10.48550/arXiv.2407.09517> // Электронный ресурс. URL: <https://arxiv.org/abs/2407.09517> (дата обращения 22.05.2025)
58. Tegmark, M. *Life 3.0: Being human in the age of artificial intelligence*. Alfred A. Knopf, 2017
59. Thompson, E. *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press, 2007
60. Tononi, G. An information integration theory of consciousness. *BMC Neuroscience*, 5(42), 2004 // Электронный ресурс, URL: <https://doi.org/10.1186/1471-2202-5-42>, дата обращения 8.06.2025
61. Tomasello M. *A natural history of human thinking*. Harvard University Press, 2014



62. Tomasello M. The Cultural Origins Of Human Cognition, London, Harvard University Press, 1999
63. Turing, A. M. Computing machinery and intelligence. 1950, Mind, 59, 433–460. // Электронный ресурс. URL: <https://doi.org/10.1093/mind/LIX.236.433> (Дата обращения 08.03.2025);
64. Turing A.M. On Computable Numbers, with an Application to the Entscheidungsproblem, 1936 // Электронный ресурс. URL: https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf (Дата обращения 08.03.2025)
65. Varela, F. J., Thompson, E., Rosch, E. The embodied mind: Cognitive science and human experience. MIT Press, 1991
66. Wolfson, I. Suffering Toasters - A New Self-Awareness Test for AI // Электронный ресурс. URL: <https://arxiv.org/pdf/2306.17258> (Дата обращения 08.03.2025)

