Философия науки и техники
2022. Т. 27. № 2. С. 136–149
УДК 004.8

Philosophy of Science and Technology
2022, vol. 27, no. 2, pp. 136–149
DOI: 10.21146/2413-9084-2022-27-2-136-149

ЯЗЫК, СОЗНАНИЕ, КОММУНИКАЦИЯ

Д.В. Винник

О некоторых вопросах архитектуры искусственной личности

Винник Дмитрий Владимирович – доктор философских наук, профессор Департамента гуманитарных наук. Финансовый университет при Правительстве Российской Федерации. Российская Федерация, 125993, г. Москва, Ленинградский пр., д. 49; e-mail: dvvinnik@fa.ru

В статье обсуждается возможность создания искусственной личности (ИЛ) как модели человеческой психики. Утверждается, что для создания искусственной личности радикальный коннекционизм неприемлем – мозг есть гибридная система, которая, помимо механизмов самообучения, содержит алгоритмические процедуры. Предлагается гипотеза функциональной роли феномена осознания самосознания как рефлексивного ранга, с которого начинается подлинная разумность. ИЛ как модель может быть построена как гибридная мультиагентная система, сочетающая цифровые и аналоговые модули, алгоритмические и самообучающиеся функции. Успехи перцептроники свидетельствуют в пользу того, что уровень моделирования ощущений может быть самообучающимся. Уровень эмоций может быть реализован в форме первичных репрезентаций в аналоговой форме с возможностью их обобщений в форме суждений как вторичных репрезентаций. Уровень самосознания или метапсихологических состояний подразумевает некую возможность обучения самообучению (как аналога осознания самосознания) с использованием на входе вторичных репрезентаций.

Ключевые слова: искусственный интеллект, коннекционизм, когнитивные функции, эмерджентизм, перцептрон, эмоции, разум, рефлексия, интенциональность, самосознание, сознание

Критика искусственной личности

Понятие сильного искусственного интеллекта (ИИ или «искин») означает возможность создания искусственной личности (ИЛ), обладающей всеми полноценными психическими атрибутами: способностью к восприятию, экспрессивному оцениванию (эмоциям), суждению, вплоть до самосознания и творческого мышления. Иными словами, понятия сильного искина и искусственной

личности можно рассматривать как тождественные. Споры о возможности конструирования личности ведутся с привлечением аргументов из самых разнообразных дисциплин, начиная от респектабельной эволюционной биологии [Shulman, Bostrom, 2012] и заканчивая сомнительной метафизикой экстерналистского толка [Manzotti, 2018].

Следует отдавать отчет, что, даже если концепция «механизма» или функционализм как его наиболее известная форма [Целищев, 2021, с. 9] истинны, а ментализм ложен, создание сильного искина в форме человекоподобного робота в обозримой перспективе следует оценивать как крайне сомнительное событие. Подразумевается, что такой робот должен иметь подлинное самосознание и, следовательно, – личность (как ту сущность, что обладает этим свойством). Эта перспектива сомнительна как с точки зрения алгоритмического подхода, так и противоположного ему коннекционизма.

С точки зрения алгоритмического подхода мы наталкиваемся на фундаментальные проблемы воспроизводства т.н. «неконцептуального содержания» феноменальной природы сознания некими формальными средствами [Crane, 1998, web]. Кроме того, нет достаточных знаний о физических ограничениях, накладываемых элементной базой мозга на реализацию тех или иных свойств, которые считаются вычислительными. Имеют ли значение для работы мозга эффекты: 1) квантового туннелирования на клеточных мембранах [Plenio, Huelga, 2008, web]; 2) «эфаптической», т.е. полевой несинаптической передачи информации [Arvanitaki, 1942, web, p. 108]; 3) устойчивой квантовой когерентности биомолекул [Chenu, Scholes, 2015]; 4) нейронных лавин как механизма амплификации квантовых эффектов [Beggs, web, 2007, p. 1344] до макроуровня и в конечном счете – до поведения [Koch, Hepp, 2006, p. 611]. Если хотя бы на 1-й и 4-й вопрос ответы окажутся положительными, можно будет утверждать, что т.н. гипотеза квантового мозга верифицирована и теории параллельных вычислений для полноценного объяснения работы мозга недостаточно. Это будет означать, что функционализм ложен. Если это так, нельзя исключать, что феноменальные свойства сознания являются не вычислительным феноменом или эпифеноменом, а неким вполне конкретным квантовым макро-феноменом функционирования нервной ткани (по аналогии с макро-феноменами сверхтекучести, сверхпроводимости и когеренции) [Винник, 2020].

С точки зрения радикального коннекционизма (концепции, что мозг есть не более чем семантически бессмысленная статистическая машина) перспективы создания ИЛ представляются более оптимистичными: если человеческий мозг есть нейронная сеть, то любая вычислительная структура, сопоставимая с ним по сложности (количеству логических вентилей и связей между ними), с необходимостью будет являться личностью. Субстрат не имеет принципиального значения – это может быть конкретный суперкомпьютер, кластер или глобальная компьютерная сеть как целое. Бытуют многочисленные спекуляции, что сама глобальная компьютерная сеть может усложниться настолько, что в ней зародится новое качество – сознание.

Допустим, что эмерджентизм (концепция спонтанного зарождения нового качества как результата эволюционного скачка сложности субстрата) истинен

и компьютерная сеть может в какой-то момент стать сопоставимой с мозгом живого существа по уровню комбинаторной сложности. Достаточное ли это основание для проявления принципиально нового качества – ментальности? Наверняка, с точки зрения приверженцев идеи нейросоциума или церебрально-открытого общества [Эпштейн, 2003, с. 358], все обстоит еще лучше – на элементном уровне разум уже встроен в глобальную сеть в виде агентов-пользователей. М. Эпштейн на данный аспект не обращает внимания, но с этой точки зрения глобальная сеть является гибридной системой из логических вентилей разной природы. В таком виде она на порядки мощнее и ближе к эмерджентному порогу феномена самосознания, чем в чистом транзисторно-проводном представлении.

Даже если между естественным разумом и искусственным интеллектом нет онтологической пропасти, из этого не следует, что техническое воспроизводство такой сокровенной сущности как разум является задачей конструктивного типа. Тому есть несколько причин, но одна из них представляется самой значительной. Дело в том, что любая естественная личность, человека или животного, является результатом сотен миллионов лет эволюции нервной системы. Те комбинации психических свойств, которые представляют единства, именуемые нами личностями, есть результат калибровки параметров торможения/возбуждения миллиардов нейронов живых существ в триллионах итераций естественного отбора.

Даже если радикальный коннекционизм ложен, отрицать существенное эволюционное значение самообучения нервных систем нельзя. Те формы сознания, которые нам известны, есть всего лишь одна из комбинаций в многомерном универсуме психических способностей. Иными словами, далеко не все сопоставимые по сложности с мозгом вычислительные системы обладают сознанием и являются личностями. Более того, может оказаться, что личностями являются не все системы, обладающие сознанием. Здравый смысл и классическая философская традиция склоняют нас рассматривать сознание именно в качестве атрибута личности, а не как самостоятельную сущность любых сложных вычислительных систем. Однако ничто не обязывает нас рассуждать именно так. Например, еще Э. Гуссерль утверждал, что, «безусловно мыслимо и бесплотное, а также, сколь парадоксально это бы ни звучало, и бездушное, не одушевляющее человеческую телесность сознание, т.е. такой поток переживания, в котором не конституировались бы интенциональные единства опытного постижения – тело, душа, эмпирический “я”-субъект». Для всех этих понятий опыта «не было бы никакой опоры, они бы были лишены всяческой значимости» [Гуссерль, 1999, с. 123].

Самонадеянно полагаться на то, что в обозримое время человеческие вычислительные ресурсы будут способны успешно воспроизвести веса (пороги возбуждения) искусственных нейронов, необходимых для самой простой модели самой примитивной личности. Способности к эмуляции эволюции и сам процесс эволюции пока несопоставимы по масштабам. Моделирование отдельных когнитивных функций, вроде распознавания образов, оказалось крайне успешным, однако этот факт часто переоценивается. Когнитивных функций великое множество, их природа различна. Кроме того, разница между

количеством искусственных нейронов количеством нейронов в мозге составляет много порядков. Сколько нейронов минимально необходимо для существования самой примитивной личности? Наверняка счет начинается как минимум с шестого порядка. Таков порядок ганглий у насекомых. Для искусственных нейронов это число должно быть еще больше, поскольку у такого нейрона есть только один тип состояния (вес от 0 до 1, моделирующий спайк), а у живых нейронов – множество дополнительных, которые обуславливают сам спайк. Кстати, по характеру спайка до сих пор нет полного согласия, является мозг цифровой или все-таки аналоговой машиной [Tee, Taylor, 2020, p. 199].

Тех, кто отрицает принципиальную возможность создания робота с подлинным самосознанием или отказывает возможности эмерджденции разума в сложных кибернетических системах как реальной, как правило, относят к сторонникам слабого искина. Последние обычно допускают отчуждение когнитивных функций без полного проникновения в сущность того, что мы считаем разумностью и человечностью. Однако это не означает, что моделирование личности как некой целостности когнитивных функций не имеет смысла.

А.Ю. Алексеев выделяет 4 разных понятия ИЛ: 1) имитация; 2) модель; 3) репродукция естественной человеческой личности; 4) креация – создание «сверхличности» [Алексеев, 2014, с. 157]. Относительно второго понятия автор пишет, что в рамках этого «когнитивно-модульного подхода» компьютерная система должна включать в свой состав блок «псевдосознания» [Там же, с. 159]. Именно о некоторых проблемных аспектах подобного программно-аппаратного моделирования личности как некоего антропоподобного единства когнитивных функций будет идти речь в настоящей статье. Ниже будут рассмотрены аспекты моделирования таких когнитивных способностей, как ощущения, эмоции, рефлексии, осознания самосознания, соответствующие разным уровням психической организации.

Ощущения и пригодность феноменологического знания

Может ли феноменологическое знание о содержании ощущений быть пригодным для задач конструирования искусственной личности? Этот вопрос полемичен. С точки зрения радикальных коннекционистов и эмерджентистов, это знание бесполезно, поскольку они считают, что структура информации на выходе системы, будь то поведение или иные ментальные феномены, есть результат калибровки параметров нейронной сети. Знание об интерфейсе ничего не дает для построения модели самой нейросетевой самообучающейся вычислительной системы: осмысленная информация может быть только на входе и на выходе, а внутри – статистический хаос весов или любых других аппроксимирующих параметров.

Сейчас очевидно, что для объяснения работы естественного интеллекта радикальный коннекционизм не годится – мозг есть гибридная система, которая, помимо механизмов самообучения, содержит алгоритмические процедуры, реализуемые разными слоями. Например, так устроена зрительная

кора – она представляет собой многослойный перцептрон, в котором каждый слой исполняет известную функцию. Базовые слои распознают геометрические примитивы: линии и углы. Более высокие слои конструируют сложные и абстрактные образы как геометрические фигуры: овалы, прямоугольники и треугольники. На верхних уровнях распознаются сложные образы и человеческие лица. Характерно, что реакции более высоких уровней менее зависят от точки зрения, реагируют на более широкую область зрительного поля и более устойчивы к искажениям. Именно это воспроизводство архитектуры зрительной коры позволило К. Фукусиме в 1980 г. создать «когнитрон» – нейросеть для распознания образов на основе конкурентного обучения, т.е. «без учителя» [Fukushima, 1988].

Вообще говоря, исследования нейрофизиологии зрения предоставили немало аргументов в пользу когнитивистов (сторонников алгоритмического понимания ментальных состояний) и против коннекционистов. П. Бреслов и Дж. Кован обратили внимание на типичность геометрических галлюцинаций – под воздействием надавливания на глаза и психофармакологических средств люди обычно видят совершенно характерные геометрические фигуры: многогранники, логарифмические спирали, матричные текстуры, решетки и туннели. Авторы утверждают, что им удалось выявить в зрительной коре те устойчивые генераторы сигналов, которые продуцируют настоящие изображения. Эти сигналы имеют вполне понятную математическую форму. Судя по всему, сигналы работают постоянно в фоновом режиме, выполняя, например, функцию опорной частоты. При определенных обстоятельствах этот скрытый уровень синтеза изображений становится доступен сознанию [Bressloff, Cowan et al., 2001, p. 299].

Известен случай американца Джейсона Паджетта, который получил тяжелую контузию. После травмы зрительное восприятие Паджетта кардинально изменилось: предметы распадались на фрагменты и только движущиеся объекты позволяли сложить целостные визуальные образы и ориентироваться в пространстве. Окружности он воспринимал исключительно как многоугольники. Струи воды, облака, лужи, радуга – для него все это состояло из крупных пикселей. Особенно сложные структуры он видел на границах раздела сред, например на краях облаков, подсвеченных солнцем. Паджетт начал рисовать структуры с помощью линейки и циркуля, чем успешно занимается уже три десятка лет. Как выяснилось, он изображал известные фракталы. Как предполагает Б. Брогард, сознанию Паджетта стали доступны процессы обработки зрительной информации нижних слоев зрительной коры. Вероятно, он стал видеть некую опорную структуру зрительного поля. Обратим внимание, что она оказалась не хаотической, а высокоупорядоченной, вписывающейся в известные и понятные математические представления [Brogaard, 2011, web].

Можно сделать вывод, что, несмотря на успехи перцептроники, феноменологические классификации и психофизическое знание в целом имеет большое значение, поскольку чувственное восприятие имеет сложную структуру: подразделяется на множество модальностей и вступает в разные отношения с ментальными состояниями других типов и модальностей.

Эмоции как первичные аналоговые репрезентации

Аристотелевская тринитарная структура души в несколько видоизмененном виде существует в общей психологии до сих пор, она включает сенсорно-перцептивный, эмоционально-экспрессивный и сознательно-волевой уровни психической организации.

Как можно наблюдать, в течение веков имеет место полемика по поводу границ между тремя уровнями психической организации. Особенно полемичен статус эмоциональных состояний – их относят то к перцептивному уровню, то к уровню самосознания. В качестве общего признака выступает их логическая форма.

Одни обращают внимание на то, что перцептивные и эмоциональные состояния являются качественными в том смысле, что их содержание в суждениях выражается предикатами первого порядка. Например, краснота и раздраженность суть простые качества.

Другие настаивают, что эмоции, как правило, являются не первичными чувственными данными, но встроены в интенциональные состояния, например в страхи и желания. Последние считаются аналитическими философами каноническими примерами интенциональных состояний. Логическая форма интенциональных состояний такова, что их содержание не может быть выражено одноместными предикатами, – это высказывания об отношении к конкретным объектам, так и к различному содержанию (выражающемуся в отдельном суждении). Важно, что отношение может различаться модальностями: как простыми поведенческими интенциями (стремление и избегание как корреляты желания и страха), так и более сложными эпистемическими (вера, убеждение, сомнение, знание). Очевидно, что такие базовые эмоции, как удовольствие и неудовольствие, сопровождают состояния желания и страха и нередко напрямую отождествляются с ними.

Подобную аргументацию можно встретить у Д. Юма, который сводил классы ментальных состояний к двум: впечатлениям (т.е. ощущениям) и идеям. Эмоции Юм причислял к содержанию рефлексивных актов: «идея удовольствия или страдания, возвращаясь в душу, производит новые впечатления – желание и отвращение, надежду и страх, которые, собственно, могут быть названы *впечатлениями рефлексии*, так как извлечены из последней» [Юм, 1996, с. 64]. Юм утверждал, что исследование наших ощущений касается скорее анатомов и естественников, чем моралистов, а «аффекты, желания и эмоции возникают по большей части из идей» [Там же, с. 65].

Достижение Юма заключается в том, что он показал, что эмоциональные состояния есть форма некой оценки содержания, безотносительно его природы – перцептивная она или абстрактная. Какова эта форма? Сейчас ясно, что эта форма имеет аналоговую природу, т.е. она измеряет интенсивность некоего параметра с помощью непрерывной шкалы. Отдадим должное факту, что известны паталогические формы или состояния психики, которые сопровождаются не только отчуждением (человек осознает некие эмоции, но не воспринимает их не как свои), но и катастрофическим вырождением эмоциональных состояний. Такой формой является состояние деперсонализации [Каплан, Сэдок, 1998, с. 430].

Создатель квантовой электродинамики Ф. Дайсон настаивает на том, что мозг – аналоговая машина, поскольку информация эмоций и понимания представлена в аналоговом виде [Dyson, 2014].

Можно сделать вывод, что, если в качестве критерия общности между эмоциями и прочими ментальными состояниями выбирается их содержание, эмоции относят к классу качественных состояний. Если в качестве критерия выбрать модальность акта, то их можно отнести к интенциональным. Так или иначе, эмоции следует рассматривать как некий контур первичной оценки ментального содержания. Судя по всему, этот контур может быть реализован на аналоговой основе.

Психофизическая природа рассудка и рефлексии

Именно модель естественной личности есть искусственная личность в собственном смысле слова. В рамках модельного подхода А.Ю. Алексеев приводит «типовую» трехуровневую архитектуру когнитивно-компьютерной системы: 1) уровень коннекционистских образов (паттернов), осуществляющий перцептивную обработку данных; 2) уровень первичных репрезентаций, переводящий восприятия в дискретные представления и суждения; 3) уровень вторичных репрезентаций, на котором осуществляется представление представлений (моделирование других моделей представления знаний и моделирование собственной модели) [Алексеев, 2008, web].

По сути, нам представлена вполне узнаваемая иерархия из перцептивного, первичного обобщающего и оценочного и рефлексивного (метапсихологического) уровней психической организации. Существуют ли подходы к моделированию рефлексивного контура психики? Обычно его описывают как моделирование или воспроизводство функции самосознания или разумности. Для этих целей используются самореферентные модели и рекурсивные функции. Речь об этом пойдет в заключительном разделе.

Как известно, разум – понятие возвышенное и метафизичеки нагруженное. Его более приземленный образ и аналог, известный как рассудок, легче подлежит формализации и моделированию. В конечном счете его можно свести к способности к суждению. Существует теория, основанная на исследовании нейрофизиологии и поведении пчел, согласно которой большого количества вычислительных ресурсов для рассудка не требуется [Chittka, Niven, 2009]. Поставив задачу численно оценить разницу между насекомыми и млекопитающими, Л. Читтка и Дж. Нивен пришли к выводу, что длина сложных последовательностей действий у млекопитающих всего втрое больше, чем у пчел. Разница же в количестве нейронов составляет 4 порядка.

Авторы приходят к выводу, что вычислительных мощностей для обслуживания когнитивных способностей необходимо гораздо меньше, чем считалось ранее [Ibid., p. 1007]. Рост нервной ткани в процессе эволюции был обусловлен в первую очередь не растущими потребностями интеллекта, а потребностями управления моторикой растущей мышечной массы и улучшения разрешающих способностей органов чувств. Например, известно, что такая специфическая способность, как распознавание лиц, имеет выделенные под

эти задачи нейрофизиологические структуры, занимающие значительный объем мозга. Важно иметь в виду, что эти структуры не совпадают со общими зрительными структурами, ответственными за распознание образов вообще.

Известно, что макаки распознают лица значительно хуже людей. Установлено, что у человека, в отличие от макак, есть вентральный затылочно-височный контур (VOT), избирательно участвующий в распознавании лиц. Функциональные МРТ-исследования выявили обширную сеть структур, вовлеченных в распознание лиц, в затылочно-височных областях, доминирующих в правом полушарии [Rossion, 2019, p. 345]. Если Л. Читтка и Дж. Нивен правы, может оказаться, что и та сущность, которую мы называем рассудком и которая должна сопровождаться сознанием, является древним и достаточно простым церебральным модулем.

Г. Нортхоф, А.М. Хентцель и др. утверждают, что самореферентные процессы опосредуются срединными структурами коры головного мозга: «Поскольку они густо и обоюдно связаны с субкортикальными срединными зонами, мы защищаем точку зрения, что интегрированная система срединных структур лежит в основе человеческой личности. Мы делаем вывод, что самореферентные процессы в срединных структурах коры головного мозга (CMS) конституируют ядро нашей личности и являются критическими для выработки чувственных переживаний личности» [Northoff, Heinzel et al., 2006, web].

К. Филиппи, Д. Финштейн и др. полагают, что самосознание является «диффузным» когнитивным процессом, пронизывающим различные слои за пределами коры [Philippi, Feinstein, Khalsa, 2012, web]. Аналогично, М. Рабинович и М. Мюезинолу пришли к выводу, что «чувство самости», «ощущения себя», которое часто используется как синоним самосознания, не имеет явной локализации: «проблема “себя” обслуживается теми когнитивными модами мозга, которые не задействованы в других когнитивных процессах. Они работают с ними в противофазе во времени. Такие моды генерируются молчащими нейронными сетями, и мы, для краткости, будем называть их “молчащими”» [Рабинович, Мюезинолу, 2010, с. 376].

Если Кант был прав и разница между рассудком и разумом носит онтологический характер, может оказаться, что субстрат этих когнитивных способностей различен. Более того, есть некие предпосылки выдвинуть гипотезу, что эти субстраты находятся на противоположных уровнях иерархии. Базовые логические функции или рассудок встроен в самое основание личности и вполне может пребывать в глубинных структурах мозга, как будет показано в следующем параграфе. Напротив, высшие когнитивные функции, т.е. самосознание и разум, «обитают» на поверхности – в неокортексе, прорастая в срединные структуры.

Обратим внимание, существует подход, согласно которому для понимания структуры природы интеллектуального субъекта понимания сущности самосознания недостаточно.

Функциональная роль осознания самосознания

Биофизик и математик Э. Танненбаум, проявивший себя в разработках в области преодоления пониженной радиозаметности, психофизической природы сна высших организмов и эволюционной динамике, утверждает, что осознание самосознания является существенным аспектом человеческой формы самосознания [Tannenbaum, 2009, p. 427]. Танненбаум связывает осознание самосознания с абстрактным мышлением и оставляет открытым вопрос о его наличии у животных. Очевидным проявлением осознания самосознания является то, что самосознающие организмы могут вывести понятие самосознания дедуктивно. Согласно этому автору, самосознание является обучаемым поведением, и оно возможно только в мозгах, которые обладают способностью к ассоциативному обучению и запоминанию. Организмы с самосознанием строят «образ себя», который, в частности, определяет узнавание себя в зеркале и успешное участие в рефлексивных играх.

На некий формальный аналог осознания самосознания в эпистемической логике, который можно рассматривать как модель, также было обращено внимание. Р. Смаллианом разработана концепция, согласно которой этот формальный аналог (“awareness of self-awareness”) можно описать как *наиболее фундаментальную форму рефлексивного мышления*, да и любого мышления.

Размышляя над геделевской проблематикой, Р. Смаллиан в главе с говорящим названием «Повышение стадий самосознания» вводит иерархию неких субъектов – «Мыслителей». Эти субъекты могут интерпретироваться в качестве математических систем и, следовательно, рефлексивных «мыслящих» автоматов. Это делается на основании стандартного инструментария пропозициональной логики и оператора «*B*», так что под *Bp* следует понимать высказывание, в которое верит Мыслитель. В то же время *Bp* является предложением, которое доказуемо в системе. Опуская подробное описание типов (рангов) Мыслителей, приведем описание 4-го типа: «Все, что вы можете доказать о Мыслителях, используя пропозициональную логику, любой Мыслитель типа 4 может доказать сам о себе, так как он знает пропозициональную логику и знает, что он Мыслитель типа 4» [Смаллиан, 2013, с. 120]. Именно это свойство Смаллиан называет «осознанием самосознания» (awareness of self-awareness) машины [Smullyan, 1987, p. 166–167]. Для любого предложения *p* Мыслитель типа 4 верит в $Bp \supset BBp$.

На значение этого доказательства в нашем понимании природы стадий самосознания и их моделирования в целях создания искусственного интеллекта обращают внимание известный математик Ю.Л. Ершов [Ершов, Целищев, 2012] и В.В. Целищев: «Система типа 4 представляет в связи “сознанием” машины главный интерес. Для более полного понимания свойств этой системы и ее взаимосвязи с фактами “сознания”, “знания” и “самосознания” представляют интерес некоторые свойства самоосознающих систем» [Целищев, 2021, с. 279].

Если Э. Танненбаум прав, что само владение понятием самосознания у людей с необходимостью говорит в пользу того, что это знание было дедуцировано из явления осознания самосознания, из этого следует, что феномен осознания самосознания играет определенную функциональную роль. Если рассуждения

Р. Смаллиана о рефлексивных машинах допускают натуралистическую интерпретацию, из этого следует, что полноценная разумность начинается на более высоких рефлексивных рангах, чем сознание и даже самосознание. Обратим внимание, что Смаллиан вводит свойства «нормальности» (если некто верит в *p*, тогда он верит в *Bp*) и «стабильности» (обратное к нормальности), которые подлежат психологическому истолкованию, впрочем, как и свойства «ненормальности» и «нестабильности», хотя последнее интроспективно трудно представимо.

Этот уровень можно назвать «нулевым» уровнем разума в том смысле, что ранги ниже осознания самосознания условно можно считать отрицательными уровнями мышления, поскольку, редуцируясь к ним, мышление теряет свою полноту [Винник, 2015]. Обратим внимание на интересную особенность. По сравнению с понятием сознания, казалось бы, более абстрактное понятие самосознания на самом деле является более конкретным по той простой причине, что оно успешно натурализуется в зеркальном тесте. Аналогично, можно надеяться, что состояние осознания самосознания также подлежит поведенческой конкретизации.

* * *

Как было указано в начале, искусственная личность как модель может быть построена как гибридная мультиагентная система. Отдельные модули и целые уровни могут быть аналоговыми, другие – цифровыми; одни системы могут самообучаться, другие – функционировать согласно алгоритмическим правилам и конкретным математическим функциям. Успешный опыт перцептроники говорит в пользу того, что уровень моделирования ощущений может быть самообучающимся. Уровень эмоций может быть реализован в форме первичных репрезентаций в аналоговой форме с возможностью их обобщений в форме суждений как вторичных репрезентаций. Уровень самосознания или метапсихологических состояний подразумевает некую возможность обучения самообучению (как аналога осознания самосознания) с использованием на входе вторичных репрезентаций.

Список литературы

Алексеев, 2008, web – *Алексеев А.Ю.* Трудности проекта искусственной личности // Искусственные общества. 2008. Т. 3. № 1. URL: https://artsoc.jes.su/s207751800000077-3-1/ (дата обращения: 22.02.2022).

Алексеев, 2014 – *Алексеев А.Ю.* Когнитотехнологические проекты искусственной личности // Гуманитарное знание и вызовы времени. М.; СПб.: Центр гуманитарных инициатив: Университетская книга, 2014. С. 156–174.

Винник, 2015 – *Винник Д.В.* Осознание самосознания как «нулевой уровень» разума // Философия науки. 2015. № 4 (67). С. 76–96.

Винник, 2020 – *Винник Д.В.* Квантовые свойства в физической организации мозга: амплификация или нивелировка? // Философия науки. 2020. № 1 (84). С. 96–118.

Гуссерль, 1999 – *Гуссерль Э.* Идеи к чистой феноменологии и феноменологической философии / Пер. с нем. А.В. Михайлова. М.: Дом интеллектуальной книги, 1999. 336 с.

Ершов, Целищев, 2012 – *Ершов Ю.Л., Целищев В.В.* Алгоритмы и вычислимость в человеческом познании. Новосибирск: Изд-во СО РАН, 2012. 504 с.

Каплан, Сэдок, 1998 – *Каплан Г.И., Сэдок Б.Дж.* Клиническая психиатрия: из синопсиса по психиатрии: в 2 т. / Пер. с англ. В.Б. Стрелец. Т. 1. М.: Медицина, 1998. 670 с.

Рабинович, Мюезинолу, 2010 – *Рабинович М.И., Мюезинолу М.К.* Нелинейная динамика мозга: эмоции и интеллектуальная деятельность // Успехи физических наук. 2010. № 4 (180). С. 371–387.

Смаллиан, 2013 – *Смаллиан Р.* Вовеки неразрешимое. Путь к Геделю через занимательные загадки / Пер. В.В. Целищева. М.: Канон+, РООИ «Реабилитация», 2013. 303 с.

Целищев, 2021 – *Целищев В.В.* Алгоритмизация мышления: Геделевский аргумент. 2-е изд., испр. М.: ЛЕНАНД, 2021. 304 с.

Эпштейн, 2003 – *Эпштейн М.* Нейросоциум // Проективный философский словарь. СПб.: Алетейя, 2003. С. 358.

Юм, 1996 – *Юм Д.* Трактат о человеческой природе, или Попытка применить основанный на опыте метод рассуждения к моральным предметам / Пер. С.И. Церетели // Сочинения: в 2 т. Т. 1. М.: Мысль, 1996. С. 53–655.

Arvanitaki, 1942, web – *Arvanitaki A.* Effects Evoked in an Axon by the Activity of a Contiguous One // Journal of Neuropsyhology. 1942. Vol. 5. No. 2. P. 89–108. URL: https://www.physiology.org/doi/abs/10.1152/jn.1942.5.2.89 (дата обращения: 22.02.2022).

Beggs, 2007, web – *Beggs M.* Neuronal avalanche // Scholarpedia. 2007. No. 2 (1). P. 1344. URL: http://www.scholarpedia.org/article/Neuronal_avalanche (дата обращения: 22.02.2022).

Bressloff, Cowan et al., 2001 – *Bressloff C., Cowan D., Golubitsky M., Thomas J., Wiener C.* Geometric Visual Hallucinations, Euclidean Symmetry and the Functional Architecture of Striate Cortex // Philosophical Transactions: Biological Sciences. 2001. No. 1407. P. 299–330.

Brogaard, 2011, web – *Brogaard B.* A Case of Acquired Savant Syndrome and Synesthesia Following a Brutal Assault. 2011. URL: https://drive.google.com/file/d/0B0GEjtSycjTKNDU4ZmVhNjktNDk2OC00MjBhLTk5ZmQtYzBhYTRkM2ZlNmU4/view (дата обращения: 22.02.2022).

Chenu, Scholes, 2015 – *Chenu A., Scholes G.D.* Coherence in energy transfer and photosynthesis // Annual Review of Physical Chemistry. 2015. No. 66. P. 69–96.

Chittka, Niven, 2009 – *Chittka L., Niven J.* Are Bigger Brains Better? // Current Biology. 2009. Vol. 19. No. 21. P. 995–1008.

Crane, 1998, web – *Crane T.* Conceptual and Non-Conceptual Content // Routledge Encyclopedia of Philosophy. Taylor and Francis, 1998. URL: https://www.rep.routledge.com/articles/thematic/content-non-conceptual/v-1 (дата обращения: 22.02.2022).

Dyson, 2014 – *Dyson F.* Are Brains Analogue or Digital? // UCD – University College Dublin. 19th May 2014. URL: https://www.youtube.com/watch?v=JLT6omWrvIw (дата обращения: 22.02.2022).

Fukushima, 1988 – *Fukushima K.* Neocognitron: A hierarchical neural network capable for visual pattern recognition // Neural networks. 1988. Vol. 1. No. 2. P. 119–130.

Kim, 1998 – *Kim J.* Philosophy of mind. Colorado: Westview Press, 1998. 352 p.

Koch, Hepp, 2006 – *Koch C., Hepp K.* Quantum mechanics in the brain // Nature. 2006. Vol. 440. P. 611. URL: https://www.nature.com/articles/440611a (дата обращения: 22.02.2022).

Manzotti, 2018 – *Manzotti R.* The Spread Mind: Why Consciousness and the World Are One Hardcover. NY, London: OR Books, 2018. 304 p.

Northoff, Heinzel et al., 2006, web – *Northoff G., Heinzel A., de Greck M., Bermpohl F., Dobrowolny H., Panksepp J.* Selfreferential processing in our brain – a meta-analysis of imaging studies on the self // NeuroImage. 2006. Vol. 31 (1). P. 440–457. URL: http://www.ncbi.nlm.nih.gov/pubmed/16466680 (дата обращения: 22.02.2022).

Philippi, Feinstein, Khalsa, 2012, web – *Philippi C., Feinstein J.S., Khalsa S.S., Damasio A., Tranel D., Landini G., Williford K., Rudrauf D.* Preserved self-awareness following extensive bilateral brain damage to the insula, anterior cingulate, and medial prefrontal cortices // PLoS ONE. 2012. Vol. 7 (8). P. e384132012. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0038413 (дата обращения: 22.02.2022).

Plenio, Huelga, 2008, web – *Plenio M.B., Huelga S.F.* Dephasing-assisted transport: quantum networks and biomolecules // New Journal of Physics. 2008. Vol. 10. URL: https://iopscience.iop.org/article/10.1088/1367-2630/10/11/113019/meta/ (дата обращения: 22.02.2022).

Rossion, 2019 – *Rossion B.* Neurophysiology of human face recognition // *Neurophysiologie Clinique*. 2019. Vol. 49 (4). P. 345.

Shulman, Bostrom, 2012 – *Shulman C., Bostrom N.* How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects // Journal of Consciousness studies. 2012. Vol. 19. No. 7–8. P. 103–130.

Smullyan, 1987 – *Smullyan R.* Forever Undecided: A Puzzle Guide to Godel. Oxford: Oxford University Press, 1987. 272 p.

Tannenbaum, 2009 – *Tannenbaum E.D.* Speculations on the emergence of self-awareness in big-brained organisms // Conscious Cognition. 2009. No. 18 (2). P. 414–427.

Tee, Taylor, 2020 – *Tee J., Taylor D.P.* Is Information in the Brain Represented in Continuous or Discrete Form? // IEEE Transactions on Molecular, Biological and Multi-Scale Communications. 2020. Vol. 6. P. 199–209.

Some aspects of artificial personality architecture

Dmitriy V. Vinnik

Financial University under the Government of Russian Federation. 49, Leningradsky av., Moscow, 125993, Russian Federation; e-mail: dvvinnik@fa.ru

This paper examines the possibility of artificial personality as a form of strong AI and as a model of the human psyche. It is argued that connectionism is not relevant for creating an artificial personality – brain is a hybrid system, which, in addition to self-learning circuits, contains algorithmic procedures. A hypothesis of the functional role of the awareness of self-consciousness property as a reflexive rank, since genuine intelligence arise, is proposed. AI as a model can be built as a hybrid multi-agent system. Some circuits may be analog, others – digital; some systems may be self-learning, others may operate under algorithmic rules and mathematical functions. The success of perceptronical inclines to idea that sensation modeling can be self-learning. The level of emotions can be realized as primary representations in the analog form. The later may be abstracted in judgments treated as secondary representations. The level of self-consciousness or meta-psychological states implies certain possibility of learning to self-learning (as an analogue of awareness of self-awareness) using secondary representations at the input.

Keywords: artificial intelligence, connectionism, cognitive functions, cognitivism, emergentism, perceptron, emotions, mind, reflection, intentionality, self-awareness, consciousness

References

Alekseev, A.Y. “Kognitotekhnologicheskie proekty iskusstvennoj lichnosti” [Cognitological progects of artificial personality], in: *Gumanitarnoe znanie i vyzovy vremeni* [Social knowledge and challengse of time]. Moskow, St. Petersburg: Centr gumanitarnyh iniciativ; Universitetskaya kniga Publ., 2014, pp. 156–174. (In Russian)

Alekseev, A.Y. “Trudnosti proekta iskusstvennoj lichnosti” [Difficalties of artificial personality project], *Iskusstvennye obshchestva* [Artificial societies]. 2008, vol. 3, no. 1 [https://artsoc.jes.su/s207751800000077-3-1/, accessed on 07.07.2022]. (In Russian)

Arvanitaki, A. “Effects Evoked in an Axon by the Activity of a Contiguous One”, *Journal of Neuropsyhology*, 1942, vol. 5, no. 2, pp. 89–108 [https://www.physiology.org/doi/abs/10.1152/jn.1942.5.2.89, accessed on 22.02.2022].

Beggs, M. “Neuronal avalanche”, *Scholarpedia,* 2007, no. 2 (1), pp. 1344 [http://www.scholarpedia.org/article/Neuronal_avalanche, accessed on 22.02.2022].

Bressloff, C., Cowan, D., Golubitsky, M., Thomas, J., Wiener, C. “Geometric Visual Hallucinations, Euclidean Symmetry and the Functional Architecture of Striate Cortex”, *Philosophical Transactions: Biological Sciences*, 2001, no. 1407, pp. 299–330.

Brogaard B. A *Case of Acquired Savant Syndrome and Synesthesia Following a Brutal Assault*. 2011 [https://drive.google.com/file/d/0B0GEjtSycjTKNDU4ZmVhNjktNDk2OC00MjBhLTk5ZmQtYzBhYTRkM2ZlNmU4/view, accessed on 22.02.2022].

Chenu, A., Scholes, G.D. “Coherence in energy transfer and photosynthesis”, *Annual Review of Physical Chemistry*, 2015, no. 66, pp. 69–96.

Chittka, L., Niven, J. “Are Bigger Brains Better?”, *Current Biology*, 2009, vol. 19, no. 21, pp. 995–1008.

Crane, T. “Conceptual and Non-Conceptual Content”, *The Routledge En-cyclopedia of Philosophy*. Taylor and Francis, 1998 [https://www.rep.routledge.com/articles/thematic/content-non-conceptual/v-1, accessed on 22.02.2022].

Dyson, F. “Are Brains Analogue or Digital?”, *UCD – University College Dublin*, 19th May 2014 [https://www.youtube.com/watch?v=JLT6omWrvIw, accessed on 22.02.2022].

Epstein, M. “Neurosocium”, in: *Proektivnyj filosofskij slovar* [Progective philosophical dictionary]. St. Petersburg: Aletejya Publ., 2003. (In Russian)

Ershov, U.L., Tselishchev, V.V. *Algoritmy i vychislimost’ v chelovecheskom poznanii* [Algorithms and Computability in Human Knowledge]. Novosibirsk: Publishing House SB RAS, 2012. 504 pp. (In Russian)

Fukushima, K. “Neocognitron: A hierarchical neural network capable for visual pattern recognition”, *Neural networks*, 1988, vol. 1, no. 2, pp. 119–130.

Hume, D. Traktat o chelovecheskoj prirode [Treatise of human nature], in: *Sochineniya v 2-h tomah* [Essays in 2 vols.], trans. by S.I. Tsereteli, vol. 1. Moskow: Mysl’ Publ., 1996, pp. 53–655. (In Russian)

Husserl, E. *Idei k chistoj fenomenologii i fenomenologicheskoj filosofii* [Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie], trans. by A.V. Mikhailov. Moscow: Dom intellektual'noj knigi Publ., 1999. 336 pp. (In Russian)

Kaplan, G.I., Sadock, B.J. *Klinicheskaya psihiatriya: iz sinopsisa po psihiatrii*: v 2-h tomah [Clinical psychiatry: from a synopsis on psychiatry: in 2 vols], vol. 1. Moskow: Medicina Publ., 1998. 670 pp. (In Russian)

Kim, J. *Philosophy of mind*. Colorado: Westview Press, 1998. 352 pp.

Koch, C., Hepp, K. “Quantum mechanics in the brain”, *Nature*, 2006, vol. 440, pp. 611 [https://www.nature.com/articles/440611a, accessed on 22.02.2022].

Manzotti, R. *The Spread Mind: Why Consciousness and the World Are One Hardcover* New Yourk, London: OR Books, 2018. 304 pp.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., Panksepp, J. “Selfreferential processing in our brain – a meta-analysis of imaging studies on the self”, *NeuroImage*, 2006, vol. 31 (1), pp. 440–457 [http://www.ncbi.nlm.nih.gov/pubmed/16466680, accessed on 22.02.2022].

Philippi, C., Feinstein, J.S., Khalsa, S.S., Damasio, A., Tranel, D., Landini, G., Williford, K., Rudrauf, D. “Preserved self-awareness following extensive bilateral brain damage to the insula,

anterior cingulate, and medial prefrontal cortices", *PLoS ONE*, 2012, vol. 7 (8), pp. e384132012 [https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0038413, accessed on 22.02.2022].

Plenio, M.B., Huelga, S.F. "Dephasing-assisted transport: quantum networks and biomolecules", *New Journal of Physics*, 2008, vol. 10 [https://iopscience.iop.org/article/10.1088/1367-2630/10/11/113019/meta/, accessed on 22.02.2022].

Rabinovich, M.I., Muezzinoglu, M.K. "Nelinejnaya dinamika mozga: emocii i intellektual'naya deyatel'nost" [Nonlinear dynamics of the brain: emotion and cognition] *Uspekhi fizicheskih nauk,* 2010, no. 4 (180), pp. 371–387. (In Russian)

Rossion, B. "Neurophysiology of human face recognition", *Neurophysiologie Clinique*. 2019, vol. 49 (4), pp. 345.

Shulman, C., Bostrom, N. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects", *Journal of Consciousness studies*, 2012, vol. 19, no. 7–8, pp. 103–130.

Smullyan, R. *Voveki nerazreshimoe. Put' k Gedelyu cherez zanimatel'nye zagadki* [Forever Undecided: A Puzzle Guide to Godel], transl. by V. Tselishev. Moskow: "Kanon+", ROOI "Reabilitatsiya" Publ., 2013. 303 pp. (In Russian)

Smullyan R. *Forever Undecided: A Puzzle Guide to Godel*. Oxford: Oxford University Press, 1987. 272 pp.

Tannenbaum, E.D. "Speculations on the emergence of self-awareness in big-brained organisms", *Conscious Cognition*, 2009, no. 18 (2), pp. 414–427.

Tee, J., Taylor, D.P. "Is Information in the Brain Represented in Continuous or Discrete Form?", *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2020, vol. 6, pp. 199–209.

Tselishchev, V.V. *Algoritmizaciya myshleniya: Gedelevskij argument* [Alghoritmization of reasoning: Gödel's Argument], 2nd edition. Moskow: LENAND Publ., 2021. 304 pp. (In Russian)

Vinnik, D.V. "Kvantovie svoystva v fisicheskoy organizacii mozga: amplificacija ili nivelirovka?" [Quantum properties in the physics of the brain: amplification or leveling?] *Filosofiya nauki,* 2020, no. 1 (84), pp. 96–118. (In Russian)

Vinnik, D.V. "Osoznanie samosoznaniya kak 'nulevoj uroven'' razuma" [Awareness of self awareness as a "zero level" of intelligence], *Filosofiya nauki,* 2015, no. 4 (67), pp. 76–96. (In Russian)