RUDN Journal of Philosophy. ISSN 2313-2302 (Print), ISSN 2408-8900 (Online) 2025 Vol. 29 No. 2 473–490
Вестник РУДН. Серия: ФИЛОСОФИЯ http://journals.rudn.ru/philosophy

https://doi.org/10.22363/2313-2302-2025-29-2-473-490
EDN: THULXV

Научная статья / Research Article

Значение философии сознания Канта для современных исследований по искусственному интеллекту

А.Г. Пушкарский

Балтийский федеральный университет имени Иммануила Канта, Калининград, Россия
pushcarskiy@mail.ru

Аннотация. С момента своего появления программа создания искусственного интеллекта опиралась на позитивистскую, антипсихологическую философскую парадигму, в которой чисто физикалистское описание процессов мышления предполагало адекватное моделирование их с помощью релевантных задачам и целям логических машин, например, Тьюринга (60–70-е годы). Оптимистические ожидания позитивных результатов сразу столкнулись как с собственно техническими трудностями, так и со сложностями чисто концептуального характера. Однако когда появилась насущная проблема философского пересмотра базовой парадигмы ИИ, теория сознания и мышления Канта всерьез не рассматривалась и подверглась критике в 90-е годы. С 2000-х годов мы видим впечатляющие успехи применения искусственных нейронных сетей с архитектурой глубокого обучения в области моделирования мышления и сложных биологических процессов. Казалось, что основная цель программы ИИ – достижения сильного ИИ, просто вопрос времени. Но непосредственная реализация концепции коннекционизма в работе с большими объемами ассоциативных и нечетких массивов информации оказалась в целом неэффективной в области представления интеллектуальных способностей сознания, особенно в репрезентации высокоуровневых знаний и точной обработке символьной информации, т.е. высших когнитивных способностей. Тогда же некоторые специалисты по ИИ и когнитивные философы обратились к философии сознания Канта, в которой была воплощена такая трансцендентальная организация макроархитектуры интеллектуальной системы, которая обладает действующей познавательной активностью, но не соответствует современным представлениям о различных механизмах обработки входных и выходных данных в когнитивной системе. Такое познание принципиально активно, поскольку оно является продуктом синтеза способности продуктивного воображения. Для выявления данной макроархитектуры применяется кантовский трансцендентальный метод, который состоит в том, что трансцендентальная архитектура любого сознания создается не в результате эмпирических исследований интеллектуальных человеческих способностей, функционирования мозговых процессов или достижений эволюционной биологии, а конструируется исходя из априорных условий самой возможности ее существования. Этот кантовский метод призван выявить априорную

структуру сознания, изоморфную любому рационально познающему субъекту. В исследовании рассматривается то, что может предложить ИИ и когнитивным наукам философия Канта.

Ключевые слова: когнитивная система, коннекционизм, представление знаний, искусственные интеллектуальные системы, философия искусственного интеллекта

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.
Финансирование. Исследование проведено при финансовой поддержке Министерства науки и высшего образования Российской Федерации, проект № 075-15-2019-1929 «Кантианская рациональность и ее потенциал в современной науке, технологиях и социальных институтах», реализуемый на базе Балтийского федерального университета имени И. Канта (Калининград).

История статьи:
Статья поступила 10.12.2024
Статья принята к публикации 07.03.2025

Для цитирования: *Пушкарский А.Г.* Значение философии сознания Канта для современных исследований по искусственному интеллекту // Вестник Российского университета дружбы народов. Серия: Философия. 2025. Т. 29. № 2. С. 473–490. https://doi.org/10.22363/2313-2302-2025-29-2-473-490

The Importance of Kant's Philosophy of Mind for Contemporary Research in Artificial Intelligence

Anatoly G. Pushkarsky

Immanuel Kant Baltic Federal University (IKBFU), Kaliningrad, Russian Federation
pushcarskiy@mail.ru

Abstract. Since its inception, the artificial intelligence program has relied on a positivistic, anti-psychological philosophical paradigm, in which a purely physicalistic description of thinking processes assumed their adequate modeling using logical machines relevant to the tasks and goals, such as Turing (1960s–70s). Optimistic expectations of positive results immediately ran into both technical difficulties and purely conceptual difficulties. However, when the urgent problem of philosophical revision of the basic AI paradigm arose, Kant's theory of consciousness and thinking was not seriously considered and was criticized in the 1990s. Since the 2000s, we have seen impressive successes in the use of artificial neural networks with deep learning architecture in the field of modeling thinking and complex biological processes. It seemed that the main goal of the AI program – achieving strong AI – was just a matter of time. But the direct implementation of the connectionism concept in working with large volumes of associative and fuzzy arrays of information turned out to be generally ineffective in the field of representing the intellectual abilities of consciousness, especially in the representation of high-level knowledge and precise processing of symbolic information, i.e. higher cognitive abilities. At the same time, some AI specialists and cognitive philosophers turned to Kant's philosophy of consciousness, which embodied such a transcendental organization of the macroarchitecture of an intellectual system that has an active cognitive activity, but does not correspond to modern ideas about the various mechanisms for processing input and output data in a cognitive system. Such cognition is fundamentally active,

since it is a product of the synthesis of the ability of productive imagination. To identify this macroarchitecture, the Kantian transcendental method is used, which consists in the fact that the transcendental architecture of any consciousness is not created as a result of empirical studies of human intellectual abilities, the functioning of brain processes or the achievements of evolutionary biology, but is constructed based on the a priori conditions of the very possibility of its existence. This Kantian method aims to reveal an a priori structure of consciousness that is isomorphic to any rationally knowing subject. The study examines what Kant's philosophy has to offer AI and cognitive science.

Keywords: cognitive system, connectionism, knowledge representation, artificial intelligent systems, philosophy of artificial intelligence

Conflict of interest. The author declares that there is no conflict of interest.
Funding of Sources. The research was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation, project No. 075-15-2019-1929 "Kantian rationality and its potential in modern science, technology and social institutions", implemented on the basis of the Immanuel Kant Baltic Federal University (Kaliningrad).

Article history:
The article was submitted on 10.12.2024
The article was accepted on 07.03.2025

For citation: Pushkarsky AG. The Importance of Kant's Philosophy of Mind for Contemporary Research in Artificial Intelligence. *RUDN Journal of Philosophy.* 2025;29(2):473–490. (In Russian). https://doi.org/10.22363/2313-2302-2025-29-2-473-490

От зарождения программы ИИ к философии Канта

Поиск более глубоких философских оснований для программы ИИ вызван тем, что изначальное представление о когнитивной структуре сознания, способной к творческому мышлению, наткнулось на непреодолимые проблемы. Вычислительная теория разума основывалась на предположении, что наше мышление – это продукт деятельности происходящих в мозге процессов. А мозг – это просто определенного рода компьютер, который с помощью закодированных строго логически формальных языков, некоторым образом представляющий ментальный язык мышления, получает способность к интеллектуальной деятельности. Такой компьютер к тому же может с огромной быстротой интерпретировать неточную информацию, поступающую, например, от сенсорных датчиков, моделирующих органы чувств, или же улавливать скрытый смысл в распознаваемой им речи. Даже не принимая откровенно функционалистскую точку зрения на сознание и мышление, всегда можно воспользоваться, по выражению Джерри Фодора, вычислительной метафорой разума. Какова бы не была природа человеческого сознания, его деятельность, или по крайней мере его интеллектуальная деятельность, может быть смоделирована на физических устройствах, реализующих, пусть и невероятно сложные, вычислительные процессы. Ведь если достигнуто более или менее адекватное понимание, как функционирует некоторая когнитивная

система, например, homo sapiens, то это позволяет формализовать ее с помощь математического аппарата, а затем реализовать полученную математическую моделью на том или ином компьютерном оборудовании. В конце XX в., компьютеры научились решать логические задачи, сочинять стихи и простые музыкальные мелодии, а также обрабатывать огромные массивы информации что имело уже практическое значение в области управления, медицины, транспорта, военного дела и т.п. Системы, созданные в рамках исследований по ИИ, и способные к выполнению отдельных интеллектуальных функций обычно свойственных человеку, изначально получили название экспертных систем или слабого искусственного интеллекта. Сегодня, когда говорят об ИИ, в большинстве случаем имеют в виду именно слабый ИИ. Те же системы, которые по своим мыслительным способностям равны или даже сильнее человеческими, называют сильным ИИ. Именно он представляет собой конечную цель программы ИИ, достижение которой даже сегодня выглядит призрачной. В современных работах по ИИ явно ощущается отсутствие уверенности достижения сильного ИИ в обозримое время. Так, в последнем издании 2022 г. обширного пособия «Искусственный интеллект: современный подход» Рассела и Норвига можно найти изложение целей и задач современной программы ИИ. Авторы уже полагают, что ИИ не будет воспроизводить человеческий интеллект и точно его моделировать. В предисловии указаны два важных аспекта: «Теперь мы больше не предполагаем, что цель фиксирована и известна системе ИИ; вместо этого система может быть не уверена в истинных целях людей, от имени которых она действует. Она должна научиться тому, что следует максимизировать, и функционировать должным образом, даже если нет уверенности в цели… Основной объединяющей темой является идея разумного агента. Мы определяем ИИ как исследование агентов, которые получают информацию из окружающей среды и выполняют действия. Каждый такой агент реализует функцию, которая сопоставляет последовательности восприятия с действиями, и мы рассматриваем различные способы представления этих функций…» [1. P. vii].

Таким образом переопределяются и уточняются задачи слабого и сильного ИИ. Хотя авторы и признают перспективы, связанные с сильным ИИ, т.е. создании в целом интеллектуальных машин, прежде всего они сосредоточены на слабом ИИ. Искусственные интеллектуальные системы могут имитировать, быть в чем-то сходным с человеческим интеллектом или даже превосходить его в конкретных отношениях, но не могут быть им в целом. В заключении они указывают: «В целом программы превосходят человеческие возможности в одних задачах и отстают в других. Единственное, чего они явно не могут сделать, – это быть в точности людьми» [1. P. 1034], и далее: «Лишь немногие исследователи ИИ обращают внимание на тест Тьюринга, предпочитая концентрироваться на производительности своих систем при решении практических задач, а не на способности имитировать людей» [1. P. 1057].

Данные утверждения особенно примечательны на фоне современных достижений коннекционизма и нейронаук в познании природы человеческого мышления и конструировании нейронных сетей с функциями глубокого обучения. Успехи искусственных нейронных сетей кроме того, что стали обыденными в нашей жизни, например, в области распознавания речи и изображений, обработки больших объемов нечеткой информации, приносят нам регулярные обескураживающие и даже тревожные сообщения, например, о случайном зарождении ИИ в нейросетях, созданных ведущими IT-компаниями, или неадекватном и пугающим поведении подобных нейросетях, или о скором вытеснении людей с большинства востребованных специальностей, которое вызовет тотальную безработицу, и т.д. и т.п.

Программа коннекционизма в самом общем смысле представляет собой направление в когнитивной науке, в котором интеллектуальные способности объясняются и моделируются с помощью искусственных нейронных сетей. Нейросети конструируются из большого числа элементов, аналогов нейронов и связей между ними, а «силы» таких связей измеряются за счет того, что им приписывается определенный вес. Такие «веса» должны моделировать работу синапсов, которые связывают между собой нейроны. Таким образом, нейросети – это упрощенные модели мозга, которые могут проявлять те или иные познавательные способности и интеллектуальные функции, сходные с человеческими, и в конце концов, как предполагается, развитие таких сетей позволит достигнуть способности к самостоятельному мышлению, аналогичному человеческому.

Хотя с момента своего появления в 1943 г., когда вышла пионерская работа Уоррена С. Мак-Каллока и Уолтера Питтса «Логическое исчисление идей, относящихся к нервной активности» [2], данное направление долго находилось на задворках исследований по ИИ. Но начиная с середины 80-х гг. XX в. многочисленные эксперименты с нейросетями продемонстрировали их нетривиальные возможности в обучении к умению в распознавании речи, образов и лиц или, например, способности к выявлению элементарных грамматических структур языков. Успехи в тех областях, которые казались непреодолимыми для стандартных цифровых компьютеров, привели к тому, что коннекционизм[1] стал восприниматься как альтернатива классической парадигме ИИ. С другой стороны, проявились и недостатки коннекционизма, особенно для программы сильного ИИ: «…хотя и коннекционистская, и классическая теории постулируют репрезентативные ментальные состояния, но только состояния, постулируемые классической теорией, связаны со знаковым уровнем ментального представления, или с “языком мысли”, то есть с репрезентативными состояниями, обладающими комбинаторной синтаксической и семантической структурой» [4. С. 230].

[1] Философское обоснование коннекционизма вызвало появление нового направления в современной философии науки – нейрофилософии. Наиболее известный представитель этого направления – Патриция Чёрчленд, см. ее книгу [3].

Джерри Алан Фодор и Зенон Вальтер Пылишин, известные специалисты по когнитивной психологии и ИИ, считают, что существуют убедительные доводы, свидетельствующие о «необходимости признать структуру разума на когнитивном уровне неконнекционистской» [4. С. 230]. Важно отметить, что оба направления исследований ИИ и когнитивным наукам изначально были явно взаимоисключающими, и каждое из них столкнулась с серьезными ограничениями, хотя и разными, в достижении главной цели ИИ – создания обобщенного ИИ. Интересно отметить, что пионеры в создании ИИ, опирающиеся на логический позитивизм и аналитическую философию, так и большинство их последователей, полагали, что философия сознания Канта вряд ли может внести ценный вклад в современную когнитивную науку и проекты ИИ. Они просто не видели в философии Канта адекватной логической теории языка, на базе которого можно построить такую «вычислительную теорию разума», которая способна к репрезентации ментальных процессов. Еще реже Кант упоминается в работах по нейронаукам и нейрофилософии. Сегодня положение меняется по причине поиска решений по выходу из тех затруднений, с которыми столкнулась программа ИИ. Показательна в этом отношении цитата калифорнийского философа Мэта Маккормика, который утверждает оригинальность философия сознания Канта и ее не редуцированность к современным функционалистским теориям сознания: «Теория мышления Канта подвергалась резкой критике в период расцвета антипсихологизма. Но я принадлежу к растущему и громогласному меньшинству философов, которые стремятся исправить мнение о том, что Кант мало что сказал правильного или полезного о философии сознания и когнитивной науке… Теперь, когда мы достигли постпозитивистского, постбихевиористского отношения к Канту, мы можем внимательнее присмотреться к функционалистскому прочтению и посмотреть, что оно и Кант могут предложить недавним попыткам смоделировать мыслящий разум» [5. P. 256].

Следует, однако, отметить, что еще в начале 90-х гг. XX в. в «Кантовском сборнике» была опубликована серия статей, посвященных проблеме применения кантовской философии в исследованиях по ИИ. В них подчеркивалось, что специалисты по ИИ, столкнувшись с указанными выше проблемами, сами стараются изобретать концепции построения когнитивных интеллектуальных систем. Между тем как в истории философии накоплены довольно значительные объемы знаний, которые могут быть использованы в разработках искусственных интеллектуальных систем: «Поэтому важной задачей для философа становится ныне перевод философских знаний о структуре и функционировании интеллекта в „процедурную“ форму, т.е. в форму, непосредственно приспособленную для включения в круг исследований по конструированию интеллектуальных систем» [6. С. 72]. Наиболее привлекательной особенностью кантовской философии, которая могла бы заинтересовать разработчиков ИИ, по мнению В.Н. Брюшинкина, будет то, что «эффективно работающая интеллектуальная система … должна сочетать в себе высокую степень

первоначальной организации с чувствительностью к опыту, т.е. должна включать в себя только такие рациональные принципы, которые заранее согласованы с опытом и могут помочь нам в его обработке и интерпретации. Именно такой пример сочетания опытной и рациональной компонент нашего знания мы встречаем в философии Канта, в которой познавательная способность человека управляется рациональными принципами, которые или обусловливают самую возможность опыта, или обладают ясными ограничениями, устраняющими их применение вне пределов возможного опыта» [6. С. 74].

Таким образом, кантовские методы построения философской системы познавательных способностей можно использовать «как образец для проектирования архитектуры ИС», т.е. интеллектуальных систем [7. С. 80], поскольку «очевидна аналогия между структурой познавательной способности и тем, что я назвал макроархитектурой систем ИИ. Разуму в таком случае соответствуют Метазнания, Рассудку – Структуры знаний об объектах и их отношениях, а Чувственности – Способы опознания объектов» [6. С. 76]. Данная схема может дать нам общий метод проектирования таких интеллектуальных систем, которые обладают «„внутренними“ возможностями расширения базисных знаний путем, в частности, рассуждений от условий возможности опыта. Эта способность к рассуждениям, основывающимся на глобальных принципах строения модели мира» [6. С. 79].

И сегодня все больше специалистов по ИИ, весьма далеких от историко-философских исследований, обратились непосредственно к текстам Канта. Ярким примером такого обращения стал вышедший в 2022 г. сборник «Кант и искусственный интеллект», где авторами наряду с философами и когнитивными психологами выступили и специалисты в области ИИ. Сборник посвящен тому, «как можно судить о требованиях и пределах искусственного интеллекта на основе кантовской философии» [8. P. vii].

Макроархитектура сознания, соотношение чувственности и рассудка и формирование модели мира

Почему же философия Канта долго оставалась вне фокуса интересов в области концептуального осмысления программы ИИ, и почему она стала актуальной в наше время хотя бы и не для большинства исследователей в данных областях. Это связано в том числе как с фундаментальными особенностями его системы, так и дальнейшим восприятием и переосмыслением ее в философии XIX и XX вв. Приведем одну цитату из книги Майкла Фридмана, стэндфордского философа, известного своей новаторской работой в кантоведении [9], которая очень показательна в этом отношении: «Кантианская система на пике Просвещения достигла замечательного синтеза практически всей человеческой мысли. Математика и естествознание, мораль и право, культура и искусство, история и религия – все это находило свое место в сложной архитектонике Канта, основанной на трех фундаментальных

способностях – чувственности, рассудке и разуме. Эти три способности, в свою очередь, описывали универсальную нормативную структуру, общую для всех людей как таковых, на все времена и во всех местах, и таким образом, поддерживали всеобщее устремление к объективности и интерсубъективной общезначимости во всех областях человеческой жизни. Это утверждение было выражено конститутивно, и так сказать, абсолютно, как в математической естественной науке, так и в морали…» [10. С. 253].

Несмотря на все достоинства кантовской философии в истории философской мысли, столь ярко выраженной Фридманом, остается открытым, действительно ли при всех различных ее интерпретациях Канту удалось раскрыть адекватную структуру когнитивной интеллектуальной деятельности и насколько они могут быть адаптированы для непосредственных технических разработок в области ИИ. Например, Ричард Эванс, ученый-компьютерщик из Deep Mind, одной из ведущих компаний, разрабатывающих современный искусственный интеллект, отмечает: «Опасность междисциплинарного проекта, с одной стороны ИИ, с другой философии, заключается в том, что обе потенциальные аудитории остаются неудовлетворенными. Информатика могла бы резонно задаться вопросом: почему книга двухсотлетней давности может нас чему-то научить сейчас? … Ученый, изучающий Канта, мог бы резонно возразить: действительно ли необходимо заново выражать теорию Канта с помощью вычислительных формализмов? … В лучшем случае это ненужная реартикуляция. В худшем случае недопонимание накапливается за недопониманием, поскольку идеи Канта неизбежно искажаются, когда их втискивают в простой вычислительный формализм» [11. P. 40]. И он же с полной определенностью дает на него ответ: «Тем не менее, я буду утверждать, что, во-первых, современному искусственному интеллекту есть чему поучиться у Канта, а во-вторых, что кантовская наука может что-то получить, если ее переформулировать на языке информатики» [11. P. 41].

Какой бы архаичной не представлялась философия Канта для исследователей нейронаук и ИИ некоторые центральные идеи, встроенные в кантовскою систему естественным образом, не могут не вызвать их интерес на современном этапе эволюции ИИ. Во-первых, у Канта мы находим описание организации макроархитектуры определенной интеллектуальной системы, принципов взаимодействия ее элементов в процессе получения и представления знаний, а также таких способов и методов их работы, которые обеспечивают именно рациональную и в современном понимании наиболее эффективную ее работу. Во-вторых, подобная макроархитектура в системе Канта в первую очередь призвана соединить эмпирические и рациональные элементы обрабатываемой информации в процессах получения и представления знаний, в которых все когнитивные способности управляется рациональными принципами, которые в свою очередь обусловливают самую возможность опыта и имеют четкие ограничения, устраняющими их применение вне пределов возможного опыта. В-третьих, по Канту любое знание о существующем мире дается нам в опыте. Однако эмпирическое знание является результатом

синтеза чувственных данных с основоположениями рассудка. Этот синтез осуществляется при помощи действий способности продуктивного воображения. А категории у Канта будут уже функциями синтеза многообразия чистого априорного созерцания. Они устанавливают связь с априорными условиями опыта и показывают невозможность их применения за пределами опыта. Исторически у Канта не было никаких актуальных данных о работе мозга, можно вполне обоснованно предположить, что лучшей моделью для синтетических процедур «кантовских» познавательных способностей могут служить нейронные сети, а при переходе к высшим когнитивным способностям, таким как логические операции с понятиями, суждениями и умозаключениям «работает» классический подход к моделированию мышления. Кратко рассмотрим только идею построения макроархитектуры сознания активно действующего субъекта, соотношение чувственности и рассудка и формирование модели мира.

Марко Беттони, один из первых ученых-компьютерщиков, которые решили обратиться к кантовскому наследию в своей деятельности, попытался определить основные условия, которым должна обладать интеллектуальная система, чтобы она была в состоянии формировать знания о мире: «Модель того способа, при помощи которого мы – посредством обработки нашего знания – создаем порядок вещей, должна удовлетворять двум основным требованиям. Это:

1. Обработка знания должна рассматриваться прежде всего, как синтетически-конституирующая (а не аналитически-трансформирующая) процедура.

2. Любая система (живая или неживая), которая сначала выполняет синтетические акты, должна моделироваться как организм, а не как машина (организм как условие синтеза)» [12. С. 135]. Если следовать Канту, то воспринимаемая нами физическая реальность не дается нам в готовом виде. Она есть продукт синтетической деятельности чувственности и рассудка, которые представляют собой совершенно разные способности познания, обладающие, в целом противоположными характеристиками. Например, чувственность пассивна и имеет дело с многообразием сенсорных данных, упорядоченных только априорными формами чувственности – временем и пространством. Рассудок – отличается активной ролью в познании, благодаря способности выносить суждения о состоянии дел и путем метафизической и трансцендентальной дедукции формировать, выводить категории, чистые и априорные понятия рассудка, и конструировать отдельные понятия мышления, т.е. придавать им значение и смысл. Объекты, как объекты мыслимой нами физической реальности, возникают только после того, когда зафиксирован пространственно-временной порядок чувственных данных в процессе познания, и под них подведено, сконструировано рассудком соответствующее понятие. В более современных терминах, можно было бы сказать, что чувственность имеет дело с аналоговой и ассоциативной информацией, а рассудок с дискретной и символьной.

В работе с первым видом информации в наше время достигнуты впечатляющие успехи в области исследований и конструирования искусственных нейронных сетей. Это стало возможным благодаря тому, что новые нейронные сети стали более приближены к организации человеческого мозга, например, к многоуровневой архитектуре зрительной коры. Их основным отличием от исторических предшественников, коннекционистских сетей из 1980-е и 1990-е гг. стало количество слоев смоделированных нейронов. В то время как классические сети состояли только из входного слоя, одного скрытого слоя и выходного слоя, глубокие нейронные сети являются глубокими в том смысле, что существует гораздо больше, чем один скрытый слой, и уже существуют такие сети, в которых количество слоев достигает несколько сотен. Это экспоненциально увеличивает вычислительную мощность нейронных сетей, что позволяет им представлять довольно абстрактные характеристики окружающей среды и в значительной степени обеспечило их современные успехи в многочисленных приложениях.

Таким образом, хотя эта новая эра коннекционизма и началась в 1980-х гг., исследователи разработали компьютеры с необходимой вычислительной мощностью только в конце 2000-х гг. Достигнутый сегодня этап развития ИИ немецкий философ Тобиас Шлихт описывает следующим образом: «После серии темных зим исследования ИИ достигли значительного прогресса благодаря появлению так называемых „архитектур глубокого обучения“ ... Этот подход машинного обучения к ИИ является одним из многих и включает в себя контролируемое, неконтролируемое обучение и обучение с подкреплением[2]. Глубокое обучение на основе искусственных нейронных сетей в настоящее время является наиболее перспективным и наиболее широко обсуждаемым (и используемым) подходом…» [13. P. 18].

Несмотря на очевидные успехи коннекционизма и нейронаук пессимисты, которых большинство среди специалистов ИИ, утверждают, что коннекционистские модели обладают определенными недостатками в представлении интеллектуальных способностей сознания, особенно

[2] В современном ИИ *машинное обучение* означает, что алгоритмы, по которым работает компьютер, имеют возможность изменятся и улучшаться при помощи обработки набора данных и выявления основных закономерностей и взаимосвязей, вскрытых в данных. *Контролируемым* обучением или обучением под присмотром или с учителем (unsupervised learning) называют процесс, когда каждому фрагменту данных присваивается метка, помогающая алгоритму машинного обучения понять значение данных. Для *неконтролируемого обучения* или обучения без присмотром или без учителя (supervised learning) обрабатываемые данные не отмечаются какими-либо описательными метками, а группируются в определенные «кластеры» на основе их сходств или различий (Например, Alpha-Go, программа, победившая чемпиона мира в классической игре Го).

Нейронные сети, которые в отличии от их первоначальных моделей, имеют очень большое число нейронов и скрытых слоев, позволяющих создавать огромное количество связей, называются в ИИ *глубоким обучением*. Так, например, система ChatGPT имеет миллиарды «нейронов». В ней реализованы такие функции как умение отвечать на вопросы, вести беседы и даже сочинять разные истории.

в репрезентации высокоуровневых знаний и точной обработке символьной информации, т.е. высших когнитивных способностей. По их мнению, основной проблемой глубоких нейронных сетей оказывается проблема понимания. Сети испытывают недостаток в богатых базовых знаниях о функциях и возможностях объектов восприятия, воспоминаниях и контекстно-зависимом познании, которые формируют человеческое познание. Как полагает американский специалист по визуальному распознаванию в системах искусственного интеллекта Мелани Митчелл, «базовые знания влияют на способность человека надежно распознавать данный объект. Даже самым успешным системам машинного зрения искусственного интеллекта не хватает такого понимания и той надежности, которую оно обеспечивает» [14. P. 132].

Изначально предполагалось, что классическая или символическая и коннекционистская парадигмы несовместимы. Развернутая аргументация в пользу такой точки зрения были выдвинуты, например, уже упомянутыми Фодором и Пылишиным [4]. Противоположная точка зрения сострит в том, что оба этих подхода не являются несовместимыми, но они возникают в результате моделирования когнитивной системы в разных масштабах или с разных точек зрения. В конце 1990-х гг. шведский философ, когнитивист Питер Гарденфорс [15] представил обоснование того, что связь между символическим и концептуальным уровнями, с одной стороны, и уровнем коннекционизма, с другой, состоит в том, что коннекционизм имеет дело с «быстрым» поведением динамической системы, в то время как концептуальные и символические структуры могут проявляться как «медленные» особенности таких система. В результате одна и та же система, в зависимости от принятой точки зрения, может рассматриваться как ассоциативный механизм и как концептуальное пространство, которое, в свою очередь, обеспечивает основу для символической системы.

Таким образом, переходя от одной точки зрения к другой, можно рассматривать концептуальные представления и символические выводы как возникающие в результате динамических процессов в системе коннекционизма. Ключевым моментом является то, что нет необходимости различать два или три типа систем, поскольку разные точки зрения могут быть приняты в одной системе обработки информации.

В последнее в области исследований по ИИ время получило развитие так называемого имплементационного подхода в коннекционизме, которое предполагает синтез двух, казалось бы, несовместимых подходов. В нем предполагается, что сама психическая деятельность, включающая обработку сенсорной информации, должна моделироваться нейронными сетями, а на более абстрактном и высоком уровне представления она уже будет реализована классической моделью мышления на основе программируемого процессора. Так Ричард Эванс отмечает, что «все больше признается, что сильные и слабые стороны нейронных сетей и обучения на основе логики дополняют друг друга…» [11. P. 41]. Он обращается к Канту, поскольку в «первой критике

Кант очень подробно описывает, как именно должна выглядеть эта гибридная архитектура. Причина, по которой он интересовался гибридными когнитивными архитектурами, заключалась в том, что он пытался синтезировать две конфликтующие философские школы того времени: эмпиризм и рационализм. Нейронная сеть – интеллектуальный предок эмпиризма, точно так же как обучение, основанное на логике, – интеллектуальный предок рационализма. Объединение эмпиризма и рационализма Кантом представляет собой когнитивную архитектуру, которая пытается объединить лучшее из обоих областей и указывает путь к гибридной архитектуре, которая сочетает в себе лучшее из нейронных сетей и подходов, основанных на логике» [11. P. 41].

Эванс и его коллеги-компьютерщики предлагают интерпретировать первую критику Канта «как точное, реализуемое с помощью вычислений описание того, как и с помощью чего осуществляется осмысление сенсорного потока» [11. P. 40]. По их мнению, можно попытаться описать когнитивную архитектуру Канта в строгой алгоритмической форме, реализовать ее на техническом уровне, а затем протестировать полученную систему экспериментально. Даже если таким путем и невозможно будет формализовать все тонкости кантовской философии сознания, точность и подробность описания на уровне компьютерного алгоритма позволит реализовать ее элементы на базе искусственной машинной системы.

Дело в том, что в первой критике Кант конструирует довольно сложную и хитроумную априорную когнитивную макроархитектуру любого сознания, которое обладало бы способностью к рациональному познанию мира. Он выделяет три основные познавательные способности такой макроархитектуры и несколько производных от них. Это чувственность (интуиция) или способность воспринимать предметы, которые непосредственно воздействуют на нас, рассудок как способность устанавливать правила и воображение. Способность воображения – это такая познавательная способность, которая отвечает за любой синтез[3], в том числе за связь между предметами чувственности и понятиями рассудка. Производной способностью, которая вытекает из рассудка, будет разум как способность выносить суждения о безусловном. Разум у Канта отвечает за нашу способность конструирования умозаключений. Еще одна высшая производная способность, связанная с рассудком – это способность суждения, которая к тому же будет как бы промежуточным между рассудком и разумом: «Если рассудок вообще провозглашается способностью устанавливать правила, то способность суждения есть умение *подводить* под правила, т. е. различать, подчинено ли нечто данному правилу (*casus datae legis*) или нет» [16. С. 187]. Она дает нам «…формальное условие, при котором нечто может быть дано в созерцании» [16. С. 323].

[3] Правда во втором издании первой критики Кант определяет еще и чистый рассудочный (или интеллектуальный) синтез, который не сводится к воображению, а как бы надстраивается над ним.

Но в дальнейшем в «Критике способности суждения» Кант расширяет и видоизменяет значение способности суждения, определяя ее как «способность мыслить особенное как подчиненное всеобщему. Если дано всеобщее (правило, принцип, закон), то способность суждения, которая подводит под него особенное (и в том случае, если она в качестве трансцендентальной способности суждения a priori указывает условия, сообразно которым только и можно подводить под это общее), есть *определяющая* способность. Но если дано только особенное, для которого надо найти всеобщее, то способность суждения есть чисто *рефлектирующая* способность» [17. С. 99].

В проекте Эванса и его коллег определенные априорные структуры сознания, которые упорядочивают и систематизируют все данные чувственного (интуитивного) познания, должны формировать шаблон системы машинного обучения, которая получит возможность перевода различных познавательных способностей и их взаимодействия в одну программу. В общем виде в их программе чувственная интуиция – это то, что обеспечивает входную информацию для когнитивной архитектуры. Рассудок в целом, как способность выносить суждения, соответствует программе с неконтролируемым обучением, а кантовская функция способности суждения, «на основании которой предмет подводится под понятие» [18. С. 403], реализована как бинарная нейронная сеть[4]. Способность воображения выполняющее функцию продуктивного синтеза, отвечающее за неизбежные взаимосвязи между интуициями, в данном проекте «реализована как набор недетерминированных правил выбора» [11. P. 95]. Эванс называет такую конструкцию «машиной апперцепции», которая «обеспечивает унифицированную реализацию различных способностей, описанных Кантом», подсистемы которой «в высшей степени недетерминированы», а «способность суждения свободна конструировать любые правил вообще – до тех пор, пока объединенный продукт трех способностей удовлетворяет различным условиям единства (реализованным как ограничения)» [11. P. 95].

Более подробно эта система описана в коллективной статье «Придание смысла исходным данным». В ней описывается «нейро-символическая основа для выделения интерпретируемых теорий из потоков исходного, необработанного сенсорного опыта». Сначала авторы расширяют «определение задачи апперцепции, включив в него неоднозначный (но все же символический) ввод: последовательности наборов дизъюнкций». Затем они используют «нейронную сеть для сопоставления необработанных сенсорных данных с дизъюнктивными входными данными», такая «бинарная нейронная сеть закодирована как логическая программа, поэтому веса сети и правила теории

[4] Бинарная нейронная сеть значительно отличается от сетей перцептронного типа и представляет собой матрицу с входами и выходами в виде наборов битов, нейроны которой реализуют функции двоичной логики нескольких переменных.

могут быть решены совместно как единая задача SAT[5]». Теперь, как утверждают авторы статьи, «мы можем совместно научиться воспринимать (сопоставлять необработанную сенсорную информацию с концепциями) и апперцепировать (объединять концепции в декларативные правила)» [19. P. 1].

Вопрос, возникающий в связи с данными разработками, состоит даже не в том, насколько будет эффективной такая «машина апперцепции» для развития ИИ. Сегодня уже несомненно, что гибридные системы будут обладать гораздо большими возможностями для решения разнообразных задач ИИ и гибкостью в применении их интеллектуального потенциала чем системы, основанные на чисто символическом или коннекционистком подходе.

Однако насколько уточнение деталей самого кантовского проекта может помочь в конструировании автономных искусственных интеллектуальных систем, способных к познанию внешнего мира и в достижении главной цели – создание сильного ИИ? Любая такая система должна иметь внутреннее представление мира, лежащее в основе формирования возможных суждений об окружающем мире, и служить базисом для принятия решений в автономно функционирующих интеллектуальных системах. Один израильский ученый в области информатики таким образом рисует современную картину состояния исследований в области ИИ: «Успехи глубокого обучения были поистине поразительными и застали многих из нас врасплох… В результате общественность считает, что „сильный ИИ“, машины, думающие как люди, уже не за горами или, возможно, даже уже здесь. На самом деле ничто не может быть дальше от истины … область искусственного интеллекта „переполнена микрооткрытиями“ – такими вещами, которые становятся хорошими пресс-релизами, – но машины по-прежнему разочаровывают и далеки от человеческого познания … Цель сильного ИИ – создать машины с интеллектом, подобным человеческому, способными общаться с людьми и направлять их. Вместо этого глубокое обучение дало нам машины с поистине впечатляющими способностями, но без интеллекта. Разница глубока и заключается в отсутствии модели реальности» [20. P. 30]. Таким образом, если нашей главной задачей является проектирование автономной интеллектуальной системы на основе философских идей Канта, то реконструкция построения действительного опыта по Канту и будет означать формирование модели представления мира для данной интеллектуальной системы.

Как должна функционировать по Канту система построения действительного опыта? Формирования опыта субъекта рассматривается Кантом в главе «О дедукции чистых рассудочных понятий» из «Аналитики понятий» первой критики как результат выполнения последовательности синтезов сознания. Однако проблема в том, что Кант не дает однозначную последовательность данных синтезов и методов их осуществления. Для того чтобы схема

[5] SAT или ВЫП – задача выполнимости булевых функций, как задача из теории вычислительной сложности состоит в том можно ли присвоить значения истинности всем переменным данной функции так чтобы она оказалась истинной.

применения кантовских синтезов могла быть использована в искусственных интеллектуальных системах, Брюшинкин предлагает реконструировать их следующим образом.

Можно предположить, что каждый компонент схемы получается из более элементарных в результате синтеза определенного типа. «Для осуществления синтезов Кант предполагает два типа способностей субъекта: апперцепцию и воображение. Трансцендентальная дедукция категорий показывает нам способ синтеза опыта из ощущений, априорных форм восприятия и категорий» [7. С. 82].

Первым видом синтеза, по его мнению, будет синтетическое единство апперцепции, которое «… создает основу для дальнейших синтезов». Синтетическое единство апперцепции «объединяет априорные созерцания и ощущения в более сложную компоненту опыта – эмпирическое созерцание» [7. С. 82].

Вторым в последовательности будет синтез схватывания, «в ходе которого из эмпирического созерцания получается восприятие» [7. С. 87]. Затем следует фигурный синтез, который связывает априорные формы чувственности с объектами эмпирического созерцания в единое целое.

Завершает последовательность синтезов трансцендентальное единство апперцепции, состоящее в построение суждений субъектно-предикатной формы, субъектами которой будут отдельные восприятия, а предикатами «выделенные в фигурном синтезе формы».

Но на этом процесс построения действительного опыта или моделей мира не исчерпывается, так как каждый акт синтеза «происходит в соответствии с категориями, а категории суть не что иное, как общие схемы отношений между явлениями (объектами)», а сама система категорий выявляет концептуальную схему искомой модели мира [7. С. 88]. Применение категорий к восприятиям раскрывается уже в кантовской «Аналитике основоположений» и одновременно выполняет две задачи, во-первых, определения границ возможного опыта и во-вторых, окончательный синтез действительного опыта. Категории применяются, с одной стороны, путем снабжения их трансцендентальными схемами, «т.е. определенными чувственными коррелятами чистых понятий рассудка» [21. С. 85], с другой определением правил применения категорий к явлениям, называемых Кантом основоположениями чистого рассудка. В завершающей статье упомянутого выше цикла по ИИ Брюшинкин делает вывод о том, что только «основоположения способности суждения завершают процедуру синтеза действительного опыта, которая согласно проведенной мною ранее аналогии может служить образцом для процедуры построения моделей мира в системах ИИ» [21. С. 89].

Заключение: от Канта к созданию сильного ИИ

На пути к сильному ИИ лежит еще одна щекотливая проблема – автономность интеллектуальной системы в принятии решений. За свободу

автономных действий субъекта в философии Канта отвечает практический разум. Его деятельность распространяется на морально-этическую сферу человека, поскольку теоретический разум ограничен исключительно получением и систематизацией научного знания и не может формулировать правила и законы морально-нравственного характера. Но генетически именно рассудок отвечает за порождение любых правил и формирование на этой основе способности выносить суждение. Поэтому любая достаточно продвинутая когнитивная система, способная к высшей интеллектуальной деятельности, неизбежно должна порождать не только правила для получения и осмысления действительного опыта, но и правила своей собственной деятельности! А поскольку любая, скажем так, этическая система заранее «встроенная» в искусственную машинную систему не может быть собственно «машинной», а будет просто нашей «человеческой», то: «… машина может сама построить теорию этики, применив этап универсализации к отдельным максимам и затем, в соответствии с полученными результатами, отобразить их в традиционные *деонтические категории* – а именно: запрещено, разрешено, обязательно» [22. P. 47]. Правда такой «сильный» ИИ, самостоятельно определяющий сам для себя правила поведения, уже невозможно будет просто так отключить или заменить целиком, и, видимо, этого-то уже и стоит серьезно опасаться.

Список литературы

[1] *Russell S., Norvig P.* Artificial Intelligence: A Modern Approach. 4th ed. Edinburgh : Pearson Education Limited, 2022.

[2] *Мак-Каллок У.С., Питтс У.* Логическое исчисление идей, относящихся к нервной активности // Автоматы / под ред. К.Э. Шеннона и Дж. Маккарти. М. : ИЛ, 1956. С. 362–384.

[3] *Churchland P.S.* Neurophilosophy: Toward a Unified Science of the Mind-Brain. Cambridge, Massachusetts : The MIT Press, 1986. DOI: 10.7551/mitpress/4952.001.0001

[4] *Фодор Дж., Пылишин З.* Коннекционизм и когнитивная структура: критический обзор // Язык и интеллект / пер. с англ. и нем., под ред. В.В. Петрова. М. : Прогресс, 1995. С. 230–313.

[5] *Mccormick M.* Questions about functionalism in Kant's philosophy of mind: lessons for cognitive science // Journal of Experimental & Theoretical Artificial Intelligence. 2003. Vol. 15. No. 2. P. 255–266. DOI: 10.1080/0952813021000055180

[6] *Брюшинкин В.Н.* «Критика чистого разума» и способы построения интеллектуальных систем // Кантовский сборник. 1989. Т. 1. № 14. С. 72–81. EDN: YUQZLF

[7] *Брюшинкин В.Н.* Кант и «искусственный интеллект»: модели мира // Кантовский сборник. 1990. Т. 1. № 15. С. 80–89. EDN: YUQZSD

[8] Kant and Artificial Intelligence / edited by H. Kim, D. Schönecker. Berlin/Boston : Walter de Gruyter GmbH, 2022.

[9] *Friedman M.* Kant and the exact sciences. Cambridge : Harvard University Press, 1992.

[10] *Фридман М.* Философия на перепутье: Карнап, Кассирер и Хайдеггер. М. : Канон+ РООИ «Реабилитация», 2021.

[11] *Evans R.* The Apperception Engine // Kant and Artificial Intelligence / edited by H. Kim, D. Schönecker. Berlin/Boston : Walter de Gruyter GmbH, 2022. P. 39–103. DOI: 10.1515/9783110706611-002
[12] *Беттони М.* Кант и кризис программного обеспечения. Предложения по построению программных систем, ориентированных на человека // Кантовский сборник. 1995. Т. 1. № 19. С. 131–137. EDN: WBASGL
[13] *Schlicht T.* Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant's Approach to Cognition // Kant and Artificial Intelligence / edited by H. Kim, D. Schönecker. Berlin/Boston : Walter de Gruyter GmbH, 2022. P. 3–38. DOI: 10.1515/9783110706611-001
[14] *Mitchell M.* Artificial Intelligence. A guide for thinking humans. London : Penguin, 2020.
[15] *Gärdenfors P.* Symbolic, Conceptual and Subconceptual Representations // Human and Machine Perception: Information Fusion. New York : Springer, 1997. P. 255–270. DOI: 10.1007/978-1-4615-5965-8_18
[16] *Кант И.* Сочинения на немецком и русском языках. Т. 2. Критика чистого разума: в 2 частях. Ч. 2 / под ред. Б. Тушлинга, Н. Мотрошиловой. М. : Наука, 2006.
[17] *Кант И.* Сочинения на немецком и русском языках. Т. 4. Критика способности суждения / под ред. Б. Тушлинга, Н. Мотрошиловой М. : Наука, 2001.
[18] *Кант И.* Сочинения на немецком и русском языках. Т. 2. Критика чистого разума: в 2 частях. Ч. 1 / под ред. Б. Тушлинга, Н. Мотрошиловой. М. : Наука, 2006.
[19] *Evans R., Bošnjak M., Buesing L., Ellis K., Pfau D., Kohli P., Sergot M*. Making sense of raw input // Artificial Intelligence. 2021. Vol. 299. Article 103521. DOI: 10.1016/j.artint.2021.103521 EDN: GOBEQM
[20] *Pearl J.* The book of Why. The new science of cause and effect. London : Penguin, 2018.
[21] *Брюшинкин В.Н.* Кант и искусственный интеллект: трансцендентальный анализ моделей мира // Кантовский сборник. 1991. Т. 1. № 16. С. 84–89. EDN: YUQZYS
[22] *Powers T.M.* Prospects for a Kantian Machine // IEEE Intelligent Systems. 2006. Vol. 21. No. 4. P. 46–51. DOI: 10.1109/MIS.2006.77

References

[1] Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 4th ed. Edinburgh: Pearson Education Limited; 2022.
[2] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. In: Shannon CE, McCarthy J, editors. *Avtomaty*. Moscow: Inostrannaya literatura publ.; 1956. (In Russian).
[3] Churchland PS. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, Massachusetts: The MIT Press; 1986. DOI: 10.7551/mitpress/4952.001.0001
[4] Fodor Dzh, Pylyshyn Z. Connectionism and cognitive architecture: A critical analysis. In: Petrov VV, editor. *Language and Intelligence*. Moscow: Progress publ.; 1995. P. 230–313. (In Russian).
[5] *Mccormick M.* Questions about functionalism in Kant's philosophy of mind: lessons for cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*. 2003;15(2):255–266. DOI: 10.1080/0952813021000055180
[6] Bryushinkin VN. "Critique of Pure Reason" and Methods of Building Intelligent Systems. *Kantian Journal.* 1989;1(14):72–81. (In Russian). EDN: YUQZLF
[7] Bryushinkin VN. Kant and "artificial intelligence": models of the world. *Kantian Journal*. 1990;1(15):80–89. (In Russian). EDN: YUQZSD
[8] Kim H, Schönecker D, editors. *Kant and Artificial Intelligence*. Berlin/Boston: Walter de Gruyter GmbH; 2022.

[9] Friedman M. *Kant and the exact sciences*. Cambridge: Harvard University Press; 1992.
[10] Friedman M. A *Parting of the Way: Carnap, Cassirer and Heidegger*. Moscow: Kanon+ publ.; 2021. (In Russian).
[11] Evans R. The Apperception Engine. In: Kim H, Schönecker D, editors. *Kant and Artificial Intelligence*. Berlin/Boston: Walter de Gruyter GmbH; 2022. P. 39–103. DOI: 10.1515/9783110706611-002
[12] Bettoni M. Kant and the Software Crisis: Proposals for Building Human-Centric Software Systems. *Kantian Journal*. 1995;1(19):131–137. (In Russian). EDN: WBASGL
[13] Schlicht T. Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant's Approach to Cognition. In: Kim H, Schönecker D, editors. *Kant and Artificial Intelligence*. Berlin/Boston: Walter de Gruyter GmbH; 2022. P. 3–38. DOI: 10.1515/9783110706611-001
[14] Mitchell M. *Artificial Intelligence. A guide for thinking humans*. London: Penguin; 2020.
[15] Gärdenfors P. Symbolic, Conceptual and Subconceptual Representations. In: *Human and Machine Perception: Information Fusion*. New York: Springer; 1997. P. 255–270. DOI: 10.1007/978-1-4615-5965-8_18
[16] Kant I. *Works in German and Russian. Vol. 2. Critique of Pure Reason: in 2 parts. Pt. 2*. Moscow: Nauka publ.; 2006.
[17] Kant I. *Works in German and Russian. Vol. 4. Critique of the Power of Judgment*. Moscow: Nauka publ.; 2001.
[18] Kant I. *Works in German and Russian. Vol. 2. Critique of Pure Reason: in 2 parts. Pt. 1*. Moscow: Nauka publ.; 2006.
[19] Evans R, Bošnjak M, Buesing L, Ellis K, Pfau D, Kohli P, et al. Making sense of raw input. *Artificial Intelligence*. 2021;299:103521. DOI: 10.1016/j.artint.2021.103521 EDN: GOBEQM
[20] Pearl J. *The book of Why. The new science of cause and effect*. London: Penguin; 2018.
[21] Bryushinkin VN. Kant and Artificial Intelligence: A Transcendental Analysis of World Models. *Kantian Journal*. 1991;1(16):84–89. (In Russian). EDN: YUQZYS
[22] Powers TM. Prospects for a Kantian Machine. *IEEE Intelligent Systems*. 2006;21(4):46–51. DOI: 10.1109/MIS.2006.77

Сведения об авторе:

Пушкарский Анатолий Геннадьевич – аналитик Академии Кантиана Высшей школы философии, истории и общественных наук, Образовательно-научный кластер «Институт образования и гуманитарных наук», Балтийский федеральный университет имени Иммануила Канта (БФУ им. И. Канта), Российская Федерация, Калининград, ул. А. Невского, д. 14. ORCID: 0000-0001-6161-3941. SPIN-код: 6885-2093. E-mail: pushcarskiy@mail.ru

About the author:

Pushkarsky Anatoly G. –Analyst at the Academia Kantiana of the Higher School of Philosophy, History and Social Sciences, Institute of Education and The Humanities Cluster "Institute of Education and Humanities", Immanuel Kant Baltic Federal University (IKBFU), 14 A. Nevskogo St., Kaliningrad, 236016, Russian Federation. ORCID: 0000-0001-6161-3941. SPIN-code: 6885-2093. E-mail: pushcarskiy@mail.ru