УДК 004.8, 141
DOI 10.25205/2541-7517-2020-18-2-30-47

Критика «Китайской комнаты» Дж. Сёрла с позиции гибридной модели построения искусственных когнитивных агентов

Р. В. Душкин

Агентство искусственного интеллекта, ООО «Дикрипто»
Москва, Россия

Аннотация

В статье рассмотрен феномен понимания смысла естественного языка и, более широко, – смысла ситуации, в которой находится когнитивный агент с учётом контекста. Дано специфическое определение понимания, находящееся на пересечении нейрофизиологии, теории информации и кибернетики. Приведена схема абстрактной архитектуры когнитивного агента произвольной природы, относительно которой утверждается, что агент с такой архитектурой может понимать в описанном в работе смысле. Также приводится критика мысленного эксперимента Дж. Сёрла «китайская комната» с позиции построения искусственных когнитивных агентов, реализованных в рамках гибридной парадигмы искусственного интеллекта. Новизна представленной работы основана на применении авторского методологического подхода к построению искусственных когнитивных агентов. В рамках этого подхода рассматривается не просто восприятие внешних стимулов из среды, а философская проблема «понимания» искусственным когнитивным агентом своих сенсорных входов. Актуальность работы вытекает из возобновившегося интереса научного сообщества к теме «сильного искусственного интеллекта». Авторский вклад в рассматриваемую тему заключается в комплексном рассмотрении с различных точек зрения темы понимания воспринимаемого искусственными когнитивными агентами с формированием предпосылок для разработки новых моделей и теории понимания в рамках искусственного интеллекта, что в перспективе сможет помочь построить целостную теорию природы человеческого разума. Статья будет интересна специалистам, работающим в области построения искусственных интеллектуальных систем и когнитивных агентов, равно как и учёным из других научных областей – в первую очередь, философии, нейрофизиологии и психологии.

Ключевые слова

философия сознания, философия искусственного интеллекта, «китайская комната», семантика, восприятие, понимание, обучение, машинное обучение, сильный искусственный интеллект

Благодарности

Автор благодарен Ю. Кочубееву за ценные идеи, полученные при обсуждении статьи

Для цитирования

Душкин Р. В. Критика «Китайской комнаты» Дж. Сёрла с позиции гибридной модели построения искусственных когнитивных агентов // Сибирский философский журнал. 2020. Т. 18, № 2. С. 30–47. DOI 10.25205/2541-7517-2020-18-2-30-47

On J. Searle's "Chinese Room" from the Hybrid Model of the Artificial Cognitive Agents Design

R. V. Dushkin

Artificial Intelligence Agency, Deecrypto LLC
Moscow, Russian Federation

Abstract

The article presents a review of the phenomenon of understanding the meaning of the natural language and, more broadly, the meaning of the situation in which the cognitive agent is located, considering the context. A specific definition of understanding is given, which is at the intersection of neurophysiology, information theory and cybernetics. The scheme of an abstract architecture of the cognitive agent (of arbitrary nature) is offered, which states that an agent with such architecture can understand in the sense described in the paper. It also provides a critique of J. Searle's mental experiment "The Chinese Room" from the point of view of the construction of artificial cognitive agents within a hybrid paradigm of artificial intelligence. The novelty of the presented work is based on the application of the author's methodological approach to the construction of artificial cognitive agents. It not only considers the perception of external stimuli from the environment, but also the philosophical problem of "understanding" by the artificial cognitive agent of its sensory inputs. The relevance of the work follows from the renewed interest of the scientific community in the theme of Strong Artificial Intelligence (or AGI). The author's contribution consists in comprehensive treatment from different points of view of the theme of understanding perceived by artificial cognitive agents. It involves the formation of prerequisites for the development of new models and the theory of understanding within the framework of artificial intelligence, which in the future will help to build a holistic theory of the nature of human mind. The article will be interesting for specialists working in the field of artificial intellectual systems and cognitive agents construction, as well as for scientists from other scientific fields – first of all, philosophy, neurophysiology and psychology.

Keywords

philosophy of mind, philosophy of artificial intelligence, Chinese room, semantics, perception, understanding, learning, machine learning, strong artificial intelligence

Acknowledgements

We are thankful to Yu. Kochyubeev for valuable ideas during the discussion of the article.

For citation

Dushkin R. V. On J. Searle's "Chinese Room" from the Hybrid Model of the Artificial Cognitive Agents Design. *Siberian Journal of Philosophy*, 2020, vol. 18, no. 2, p. 30–47. (in Russ.) DOI 10.25205/2541-7517-2020-18-2-30-47

«Китайская комната» – мысленный эксперимент, находящийся на стыке философии сознания и философии искусственного интеллекта, который был предложен в 1980 г. Джоном Сёрлом [Searle, 1980]. Возможно, что это самый обсуждаемый мысленный эксперимент из всех, что были предложены в этой области. Тем не менее, до сих пор вопрос, который поставил Дж. Сёрл относительно китайской комнаты, так и не решён – «понимает ли китайская комната китайский язык?» [Searle, 2001]. Интерес представляет то, что Дж. Сёрл в своей статье постарался сразу же дать ответы на широкий набор аргументов своих будущих оппонентов, тем не менее в последующие годы на тему «китайской комнаты» вышло поразительно большое число публикаций.

Вместе с тем, прошедшие 40 лет с момента первоначальной публикации ознаменовались мощнейшим рывком как в области теоретического осмысления теории искусственного интеллекта (ИИ), так и, в особенности, в области развития и применения прикладных ИИ-технологий [Душкин, 2019]. Достижения имеются в том числе и в понимании методов построения искусственных когнитивных агентов общего уровня (AGI – art*ificial general intelligence*, англ. искусственный интеллект общего назначения, что в терминологии Дж. Сёрла называется «Сильным ИИ»). Например, автором в работе [Душкин, Андронов, 2019] предложена модель гибридной архитектуры искусственного когнитивного агента, которая может стать прототипом AGI (и эта архитектура будет кратко рассмотрена в следующем разделе). Поэтому имеется резон рассмотреть «китайскую комнату» с этих позиций.

На взгляд автора методологически проблема с «китайской комнатой» Дж. Сёрла заключается в том, что в этом мысленном эксперименте была попытка обосновать невозможность построения сильного ИИ, в то время как в пример приводилась схема слабого ИИ-агента, основанного на вычислительном подходе в рамках архитектуры фон Неймана. Это выглядит примерно так же, как если бы в отношении человека разумного был бы задан вопрос «понимает ли какой-либо конкретный нейрон (да и хотя бы комплекс нейронов) в его неокортексе то, что этот человек в заданный момент времени читает?».

С одной стороны это отсылка к так называемому «картезианскому театру» [Dennett, 1991] – по современным воззрениям, в нервной системе человека нет «центра сознания», поэтому понимает смысл человек (его центральная нервная система) в целом, а не какой-то отдельный комплекс нейронов. С другой стороны, в описании мысленного эксперимента не было дано чёткого определения понятию «понимание», в связи с чем дискуссия о понимании «китайской комнатой» смысла воспринимаемых высказываний спустилась на уровень интуитивного подхода к искусственному интеллекту [Turing, 1950], который сегодня не рассматривается в серьёзных научных и инженерных кругах. Поэтому в настоящей работе будет

дана попытка дать операционное определение феномену понимания смысла когнитивным агентом произвольной природы.

Наконец, важным вопросом в области философии искусственного интеллекта, который напрямую следует из выводов рассматриваемого мысленного эксперимента, является следующий – смогут ли когда-нибудь искусственные когнитивные агенты (ИИ-системы) понимать смысл того, что происходит вокруг них и получаемые на их сенсорные входы сигналы. Ответ на этот вопрос зависит от понимания природы «понимания смысла», и в настоящей работе даётся обоснование тезисов о том, что искусственные когнитивные агенты имеют принципиальную возможность понимать смысл с учётом «жизненного опыта» и контекста.

Новизна представленной работы основана на применении авторского методологического подхода к построению искусственных когнитивных агентов [Душкин, Андронов, 2019], притом что в рамках этого подхода рассматривается не просто восприятие внешних стимулов из среды, а философская проблема «понимания» искусственным когнитивным агентом своих сенсорных входов. Такое понимание основано на построении описания динамической модели воспринимаемой окружающей реальности с учётом как непрерывности актов восприятия, так и применения к процессу восприятия «личного опыта» агента и контекста, в котором он находится.

Актуальность работы вытекает из возобновившегося интереса научного сообщества к теме сильного искусственного интеллекта. Авторский вклад в рассматриваемую тему заключается в комплексном рассмотрении с различных точек зрения понимания воспринимаемого искусственными когнитивными агентами с формированием предпосылок для разработки новых моделей и теории понимания в рамках искусственного интеллекта, что в перспективе сможет помочь построить целостную теорию природы человеческого разума.

Статья будет интересна специалистам, работающим в области построения искусственных интеллектуальных систем и когнитивных агентов, равно как и учёным из других научных областей – в первую очередь, философии, нейрофизиологии, психологии и т. д. Автор приглашает всех заинтересованных к междисциплинарному диалогу для выработки моделей, методов и подходов к изучению и построению как отдельных модулей и функций AGI, так и сильного ИИ в целом.

1. Кратко о гибридной архитектуре ИИ-агентов

В работе [Душкин, Андронов, 2019] дано определение и описание гибридной архитектуры искусственного когнитивного агента. Тем не менее, имеет смысл

привести эту архитектуру здесь и кратко отразить основные её свойства, а также привести мотивацию, лежащую в основе её разработки.

Авторская классификация методов искусственного интеллекта (как междисциплинарной области исследований) основана на выделении двух парадигм – нисходящей и восходящей, как это было определено Дж. Маккарти и М. Минским в лаборатории информатики и искусственного интеллекта МТИ [Душкин, 2019]. Нисходящая парадигма искусственного интеллекта объединяет логический и символьный подход к построению ИИ-агентов, которые основаны на использовании формальной логики и символьной математики. Восходящая парадигма включает в себя структурный (искусственные нейронные сети), эволюционный (генетические алгоритмы) и квазибиологический подходы (использование химических реакций для осуществления вычислений).

Интерес вызывает то, что две представленные парадигмы обладают важными свойствами. При помощи методов нисходящей парадигмы можно спроектировать и реализовать ИИ-агентов, которых сложно обучать (все знания в них описаны эксплицитно в процессе реализации), но при этом они более или менее легко могут объяснить принятые в процессе своего функционирования решения. С другой стороны, ИИ-агенты, основанные на восходящей парадигме, могут быть легко обучены – именно поэтому большая часть методов восходящей парадигмы объединяются термином «машинное обучение». Однако объяснение и интерпретация принятых решений такими ИИ-агентами крайне затруднены [Шумский, 2020].

На помощь приходит так называемая «гибридная парадигма», в рамках которой объединяются два вышеупомянутых подхода к построению искусственных когнитивных агентов, от которых берётся лучшее, при этом негативные аспекты нивелируются. Другими словами, ИИ-агенты, построенные по гибридной парадигме, могут решать сложные когнитивные задачи восходящими методами, но при этом имеется возможность использования существенных выгод нисходящих методов – моделирования, прогнозирования и объяснения принятых решений.

На следующей диаграмме показана обобщённая гибридная архитектура искусственного когнитивного агента (упрощение и переработка технической схемы из [Душкин, Андронов, 2019]).

Фактически на представленной диаграмме показана стандартная схема агента, действующего в какой-либо окружающей среде, расширенная специальными нюансами, которые пояснены ниже. Особенности гибридного подхода к проектированию и реализации искусственных когнитивных агентов заключается в одновременном использовании методов восходящей и нисходящей парадигмы искусственного интеллекта. В рассматриваемом случае это означает, что когнитивный агент применяет как нейросетевые методы анализа входной информации, так

и символьные методы принятия решений. Кроме того, во внимание принят организмический подход, когда проект искусственного когнитивного агента частично основывается на известных принципах функционирования естественных когнитивных агентов.

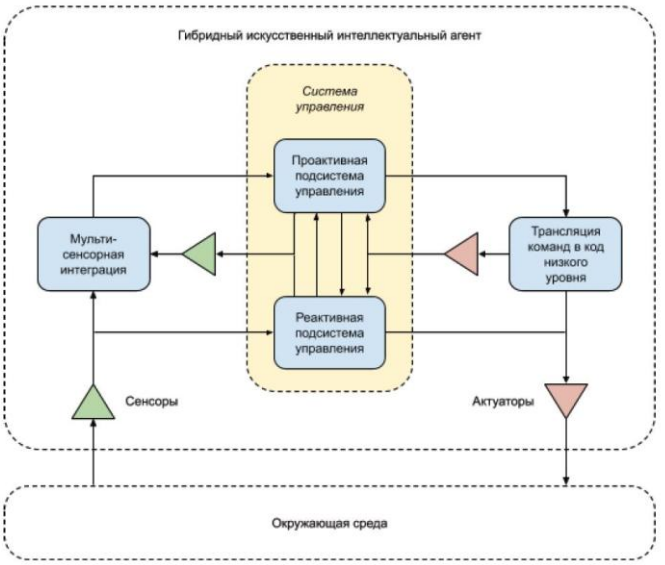

Рис. 1. Обобщённая гибридная архитектура искусственного когнитивного агента

1. В частности, вся информация из окружающей среды воспринимается агентом через сенсоры (в этом нет никаких особенностей – все агенты устроены именно так, и иное вряд ли можно себе представить). Затем сенсорная информация очищается и подаётся на вход блока сенсорной интеграции. Сюда же подаётся информация с сенсоров мониторинга внутреннего состояния когнитивного агента для обеспечения гомеостаза жизненных показателей. Модуль сенсорной интеграции как раз основан на нейросетевом подходе – здесь осуществляются базовые процессы восприятия, распознавания и когниции.

2. Вместе с тем информация от сенсоров может быть сразу подана в реактивную подсистему управления, где осуществляется поиск готовых паттернов реагирования на имеющуюся в среде ситуацию. Если такой паттерн имеется, он задействуется непосредственно для осуществления воздействия на окружающую среду. Однако проактивная подсистема управления имеет возможность подавить реак-

тивный контур, если в процессе её работы обнаружены какие-либо особенности описания среды, которые требуют более внимательного подхода к разработке реакции агента.

3. Мультисенсорная интеграция заключается в создании целостного описания ситуации внешней среды, в которой оказался агент. Такое описание подаётся на вход в проактивную подсистему управления для принятия решения о том, как агент должен действовать.

4. Проактивная система управления обладает памятью, в которой осуществляется накопление опыта агента, а также средствами моделирования поведения агента в среде и прогнозирования вариантов её состояния после актов воздействия на неё. Это позволяет осуществлять интеллектуальное планирование деятельности агента на основе решения «внутреннего конфликта» с выбором максимально эффективного варианта действия.

5. Этот выбранный вариант действия подаётся на вход модулю трансляции высокоуровневой команды на язык актуаторов, что может выражаться в последовательном программировании деятельности исполнительных устройств агента для активного взаимодействия со средой. Также здесь может осуществляться воздействие и на внутренние состояния самого агента через внутренние актуаторы.

6. Кроме того, проактивная подсистема управления спускает вновь созданную программу действий в реактивную подсистему управления для дальнейшего быстрого реагирования в случаях, когда среда находится в приблизительно таком же состоянии, в каком она находилась в то время, как проактивная подсистема управления разработала новую программу действий. В дальнейшем эта программа будет автоматически исполняться реактивной подсистемой управления.

Представленная схема деятельности гибридного искусственного когнитивного агента является достаточно обобщённой для того, чтобы конкретизировать её специфическими методами распознавания, моделирования, прогнозирования и решения всех остальных перечисленных в приведённом ранее описании задач, которые в цикле решаются представленной архитектурой. Это позволяет на основе этого шаблона проектировать конкретизированных когнитивных агентов, которые работают на основании тех или иных методов искусственного интеллекта в средах разной природы.

Также необходимо отметить, что под представленную архитектуру подходят и естественные когнитивные агенты – в частности, человек. Действительно, человек получает информацию о внешней среде через набор сенсорных систем, а также в организме осуществляется непрерывный мониторинг жизненных показателей с обеспечением их гомеостаза [Cannon, 1926]. Далее в таламусе осуществляется мультисенсорная интеграция информации для формирования в неокортексе цело-

стной картины воспринимаемой реальности [Melchitzky, Lewis, 2009]. Если ситуация требует мгновенного принятия и реализации решения, то задействуется рефлекторный контур (например, в случае, если человек дотрагивается до обжигающего объекта рукой), однако действие многих (но не всех) рефлексов может быть сознательно подавлено. В различных зонах неокортекса осуществляется базовое распознавание образов, узнавание ситуации в целом, моделирование, прогнозирование и принятие решений с дальнейшим программированием последовательности действий. После этого, если дана команда действовать, эта программа реализуется через её перевод в последовательность нейрональных активаций множества путей к мышцам для воздействия на среду [Ashby, 1960]. В процессе доведения таких команд до автоматизма сознание и сознательное размышление всё реже задействуются при решении схожих проблем, а сама программа реагирования «спускается» в мозжечок (см.: [Шмидт, Тевс, 1996. С. 107–112]).

Приведённое описание общей схемы деятельности человека лежит в основе той мотивации, которая является фундаментом для построения гибридных схем организации искусственных когнитивных агентов и, в частности, представленной в этом разделе гибридной архитектуры.

2. Что понимается под «пониманием»

Понимание – одна из важнейших концепций философии сознания в частности и философии вообще. Если расширить антропоцентрическое определение понимания, то под ним следует понимать определённый процесс, который некоторым образом соотносится с абстрактной концепцией, физическим объектом или явлением, что позволяет адекватно взаимодействовать с таковой концепцией, объектом или явлением. Понимание – это одно из достаточных свойств интеллектуального поведения [Bereiter, 2009].

Вместе с тем, в своей оригинальной работе Дж. Сёрл не дал точного определения феномену «понимания», а просто задал вопрос: «Понимает ли сидящий в китайской комнате человек китайский язык?» [Searle, 2001]. Фактически это была апелляция к интуитивному пониманию феномена понимания. И, в общем-то, тем самым Дж. Сёрл открыл широчайшие возможности для различных трактовок своего мысленного эксперимента, из-за чего «китайская комната» и стала самым дискутируемым экспериментом в истории философии сознания.

Сам Дж. Сёрл попытался опровергнуть возможность создания так называемого сильного искусственного интеллекта, т. е. набора технологий, применение которых могут привести к самосознающему искусственному когнитивному агенту, обладающему пониманием в человеческом смысле этого слова, «что бы это ни значи-

ло». Более того, у такого когнитивного агента, возможно, появилось бы и сознание, в том числе феноменальное. По словам Дж. Сёрла, «*соответствующим образом запрограммированный компьютер с нужными входами и выходами и будет разумом, в том смысле, в котором человеческий разум – это разум*» [Searle, 1980].

При этом если обратиться к нейрофизиологическим основам работы естественных нейронных сетей в режиме обучения, то тайна феномена «понимания» будет приоткрыта. В соответствии с современными воззрениями [Шумский, 2020] в процессе восприятия и обучения в неокортексе человека строится большое количество ассоциативных связей между так называемыми «неокортикальными гиперколонками», каждая из которых состоит из большого количества колонок, активирующихся на появление в рецептивных полях сенсорных зон коры головного мозга воспринимаемых образов различной модальности. Фактически активация неокортикальной гиперколонки означает, что в этот момент в потоке мыслей человека задействован образ того объекта, явления или абстрактного понятия, которому соответствует активированная гиперколонка и конкретный набор колонок в ней.

В процессе развития человека от эмбриона до взрослого состояния включительно в головном мозге непрерывно и постоянно осуществляется огромное множество актов обучения, которые заключаются в выстраивании или удалении ассоциативных связей между семантическими понятиями, что на физиологическом уровне соответствует появлению или уничтожению синаптических связей между нейронами различных кортикальных колонок [Stout, Khreisheh, 2015]. Такие синапсы позволяют осуществить активацию обширного набора смежных понятий при осмыслении какого-либо объекта или явления. И чем богаче у человека жизненный опыт, тем больше ассоциативных связей в коре головного мозга имеется, тем больше ассоциаций задействуется в процессе мышления (при этом, конечно, со стороны базальных ганглиев осуществляется «дирижирование» ансамблем активаций так, чтобы он соответствовал текущей цели) [Шумский, 2020].

Один из аспектов нейропластичности – выстраивание синаптических связей – как раз и отвечает за массовое создание ассоциаций в неокортексе человека [Chang, 2014]. Важным моментом в этом процессе является так называемая мультисенсорная интеграция, происходящая в таламусе и отвечающая за целостное восприятие окружающей действительности [Jones, 1985]. И в процессе обучения ребёнка через мультисенсорную интеграцию в его неокортексе формируется поражающее воображение количество ассоциативных связей между гиперколонками. Например, если рассмотреть понятие «стол», то ему будет соответствовать гиперколонка, в которой посредством разреженного кодирования осуществляется активация в ответ на появление перед обученным ребёнком изображений столов

различного вида, на проявление акустических волн, которые воспринимаются как звуковая форма слова «стол», и даже на появление изображений букв, складывающихся в соответствующее слово. Более того, колонки, входящие в состав этой гиперколонки, будут активировать различные смежные понятия, в том числе и общий шаблон стола, понятия ножек, поверхности, горизонтальности, гладкости, деревянности и т. д. Обучение тому, что такое «стол», и представляет собой выстраивание всех этих ассоциативных связей, которые возбуждаются в ответ на появление связанного с объектом действительности «стол» сенсорного сигнала любой модальности.

Таким образом, если у такого обученного тому, что такое «стол», ребёнка спросить, *понимает ли* он, что такое «стол», то он ответит положительно. И природа его понимания как раз и лежит в активации обозначенных связей между различными представлениями объекта реальности в его неокортексе. Другими словами, *понимание – это узнавание и активация ассоциативных взаимосвязей*. Ведь если человек что-то не понимает, это происходит из-за того, что у него в процессе когнитивной деятельности не активировалось достаточное количество ассоциативных связей с объектами, уже бывшими в его личном опыте познания, а потому объект непонимания не может встроиться в иерархическую структуру ассоциативных взаимосвязей его кортикальных гиперколонок.

Что произойдёт, если обученному таким образом ребёнку показать надпись: «桌子» (zhuōzi, кит. «стол»)? В его неокортексе не будет «отклика» ассоциативных связей, собранных в процессе получения личного опыта, а потому эти китайские иероглифы не создадут перед его внутренним взором образа «шаблонного стола». Однако если научить и закрепить изученное, то эти два символа включатся в когнитивное поле понятия «стол», и сначала через перевод на русский язык, а потом напрямую будут возбуждать все необходимые ассоциации. И, таким образом, ребёнок научится понимать китайский язык. Понимать именно в том смысле, который описан здесь несколькими абзацами ранее.

Из приведённых выше рассуждений следует интересное когнитивное искажение о понимании, которому подвержены многие люди. Дело в том, что личность человека основана на его личном опыте [LeDoux, 2020], который определяет уникальный *коннектом*, свойственный конкретному человеку. Коннектом представляет собой всё динамическое множество связей между нейронами, и из этого следует, что связи между кортикальными колонками и гиперколонками также уникальны для каждого человека. Несмотря на то что на высоких уровнях абстракции у людей формируются более или менее схожие иерархические системы понятий, на нижних конкретизированных уровнях понимание основано на личном опыте. Например, внутренняя конкретизация понятия «стол» у каждого чело-

века будет осуществляться при помощи специфических образов конкретных столов, запечатлённых в памяти человека в виде энграмм, полученных чаще всего в детстве.

Более выпуклый пример – внутреннее представление воображаемых образов во время чтения художественного произведения. Пусть некий автор описал в своём произведении какую-либо местность, в которой разворачиваются события. Наверняка при создании произведения автор представлял внутри себя какую-то конкретную местность из своего детства. При чтении же каждый читатель будет представлять ту местность, которая известна ему и только ему по его личному детскому опыту. Детскому потому, что именно в детстве осуществляется наполнение памяти очень красочными впечатлениями от воспринятого в окружающей действительности.

Другими словами, из этого следует, что каждый человек понимает смысл по-своему. И когнитивным искажением является желание считать, что все люди понимают сказанное и воспринятое одинаково. Это далеко не так. Поэтому первоначальный вопрос Дж. Сёрла, обращённый к «китайской комнате», не имеет особого смысла даже для людей.

3. Смогут ли искусственные когнитивные агенты понимать смысл

Тем не менее, если перейти от нейрофизиологических основ памяти и понимания к рассмотрению возможностей для искусственных когнитивных агентов (систем искусственного интеллекта) получить функцию понимания, так как именно на эти возможности был направлен первоначальный замысел мысленного эксперимента «китайская комната», то тут всё не так радужно. Здесь видится некоторая методологическая ошибка, которую за многочисленным преломлением копий мало кто замечает. Вместе с тем, сама постановка эксперимента даёт основания полагать, что он некорректно отвечает на вопрос о том, смогут ли когда-нибудь и как-нибудь искусственные когнитивные агенты понимать смысл воспринимаемой информации так же, как это делает человек.

Дж. Сёрл описывает некоторую аналогию фон-неймановской архитектуре построения вычислительных систем, в которой центральным процессором является человек, сидящий в комнате. И он задаёт вопрос об этом человеке, то есть, если рассматривать аналогию, – о центральном процессоре компьютера. Но у человека в центральной нервной системе нет аналога компьютерного процессора, и обработка входной информации в человеческом мозге осуществляется на совершенно иных принципах, а сама она построена на иной архитектуре. И при этом работа

нервной системы не выходит за рамки вычислительной парадигмы – она вполне может быть объяснена в терминах теории информации и кибернетики. Иными словами, человек – это тоже вычислительная система, довольно сложная, но, тем не менее, работающая на базовых принципах математики.

Вместе с тем, Дж. Сёрл задаётся вопросом о том, понимает ли представленная в мысленном эксперименте аналогия фон-неймановской архитектуры входную информацию. Ответ, по его мнению, отрицательный, и это довольно резонно им аргументировано. Однако методологическая ошибка заключается в том, что представленная в мысленном эксперименте система представляет собой так называемую слабую ИИ-систему. Слабым искусственным интеллектом называют такую ИИ-систему, которая направлена на решение какой-то конкретной задачи при помощи когнитивных методов. Например, система распознавания визуальных образов является слабой ИИ-системой. Резонно ли задавать вопрос о понимании по отношению к слабой ИИ-системе, если само по себе понимание как феномен относится к прерогативе сильных когнитивных агентов? Ответ очевиден.

Если рассмотреть современную теорию искусственного интеллекта [Душкин, 2019; Шумский, 2020], то все современные технологии решения когнитивных задач можно свести к нескольким: распознавание образов, поиск скрытых закономерностей, обработка и понимание текста на естественном языке, а также принятие решений. Слабые ИИ-агенты по определению решают какую-то специфическую задачу, то есть такая задача сводится к одной из четырёх перечисленных. Среди них только задача обработки и понимания текста на естественном языке так или иначе имеет отношение к предмету настоящей работы – пониманию. Остальные задачи коррелируют и связаны с пониманием, однако не требуют его как такового для своего решения. Поэтому слабые ИИ-системы, решающие эти задачи, по определению пониманием не обладают.

Интересно рассмотреть слабые ИИ-системы, предназначенные для решения задач анализа естественно-языковых высказываний. Задача понимания естественного языка в рамках искусственного интеллекта поставлена – *natural language understanding*, NLU. Тем не менее, на текущий момент большинство существующих технологий решают задачу обработки естественного языка – *natural language processing*, NLP. Это подразумевает отсутствие необходимости понимания в том смысле, который описан в предыдущем разделе. Текущие технологии NLP, основанные как на формальных грамматиках, так и на статистическом и нейросетевом подходах, фактически показывают реактивную модель искусственных когнитивных агентов, когда суждения о качестве их функционирования и решения ими поставленной задачи делаются на основе наблюдения за их внешним поведением. Если искусственный когнитивный агент адекватно реагирует на большую часть

обращённых к нему естественно-языковых высказываний, значит, он «понимает» их.

Но, конечно, такие же вопросы можно обратить к человеку или даже другим высшим животным. Действительно, понимает ли, например, собака? Если изучать её поведение, то вне всяких сомнений после некоторого количества обучающих актов собака начинает понимать обращённую к ней речь. «Пойдём гулять» – говорит хозяин, и собака радостно прыгает, виляя хвостом. И речь, скорее всего, не идёт о примитивных рефлекторных реакциях, так как в процессе узнавания ключевых слов в составе обращённых к ней фраз, собака также принимает во внимание внутреннее состояние, выражающееся в эмоциональном фоне и, как минимум, в наличии памяти о недавно прошедших событиях.

Однако можно предположить на основе сравнения строения нервной системы собаки и человека, что в какой-то мере все рассуждения из области нейрофизиологии восприятия, узнавания и понимания, приведённые в предыдущем разделе, относятся и к собаке, а также к другим высшим животным и даже птицам, хотя у них нет коры головного мозга. В нервной системе животных существуют механизмы массового создания ассоциативных связей между группами нейронов, отвечающих за распознавание сенсорных образов или узнавание абстрактных понятий. Так что в неокортексе собаки из слуховой сенсорной коры создаются ассоциативные связи от нейронов, распознающих фразу «Пойдём гулять», направленные в области памяти и положительных эмоций, связанных с прогулкой, и далее к моторным зонам с программами выражения радости. И это действительно можно назвать пониманием, так как механизм аналогичен человеческому.

Но как можно было бы спроектировать понимающего ИИ-агента? (Вопрос о том, нужно ли это, следует оставить за рамками этой работы, поскольку здесь делается попытка дать ответ на вопрос Дж. Сёрла по поводу понимания.) На текущий момент имеется как минимум один пример естественного агента с более или менее понятной архитектурой, про которого можно сказать, что он «умеет понимать» в том смысле, который был дан этому термину в настоящей работе. На этой аналогии можно построить проект понимающего ИИ-агента, который смог бы стать предтечей ИИ общего уровня (AGI). Да, на это желание можно было бы возразить, что «самолёт летает не так, как птица, а всё же летает», а потому понимающий ИИ-агент не обязательно должен копировать принципы, лежащие в основе человеческого понимания. Тем не менее, необходимо с чего-то начать, а дальше уже пытаться рассматривать обобщающие концепции, в том числе и при помощи рассмотрения когнитивных способностей птиц, особенно в части понимания.

В первом разделе настоящей статьи была приведена архитектура гибридного искусственного когнитивного агента, на базе которой можно было бы попытаться

реализовать функцию понимания. Действительно, естественные когнитивные агенты (люди, собаки, попугаи и т. д.) основаны именно на этой архитектуре. Поэтому в соответствии с принципом, провозглашённым в предыдущем абзаце, именно эта архитектура должна быть взята за основу для понимающего искусственного когнитивного агента.

Таким образом, проектирование и разработка искусственного интеллектуального агента, который мог бы обладать пониманием, может осуществляться на следующих принципах.

1. ИИ-агент должен иметь набор разнообразных сенсоров, при помощи которых он взаимодействует с окружающей средой и получает из неё сенсорную информацию различной модальности, а также сенсорную информацию о внутреннем состоянии самого ИИ-агента. Важно иметь сенсоры нескольких модальностей, как минимум двух, но лучше больше. Впрочем, вопрос о верхней границе количества сенсорных модальностей открыт и требует дополнительного исследования.

2. Сенсорная информация, приходящая из среды, должна подвергаться фильтрации и агрегированию непосредственно в сенсорных системах ИИ-агента. Затем очищенная и агрегированная информация передаётся далее в реактивную подсистему управления и в блок мультисенсорной интеграции.

3. Реактивная подсистема управления ИИ-агента представляет собой систему, которая осуществляет быструю рефлекторную реакцию на известные стимулы. Однако работа этой подсистемы может быть подавлена сигналом из проактивной подсистемы управления, которая предназначена для моделирования будущего состояния среды и ИИ-агента в ней для более вдумчивого принятия решения. Кроме того, реактивная подсистема управления может эскалировать в проактивную фокус внимания в случаях, когда что-то пошло не так – в известной ситуации, требующей рефлекторной реакции, среда отвечает не так, как обычно.

4. Кроме того, сигналы из сенсоров ИИ-агента поступают в блок мультисенсорной интеграции, где должно осуществляться построение целостной картины окружающей среды на основе технологии слияния данных. Именно здесь формируется тот самый набор ассоциативных связей, который представляет собой память ИИ-агента обо всём, что с ним происходило – его личный опыт функционирования.

5. Память ИИ-агента, располагающаяся в проактивной подсистеме управления, должна быть устроена на иерархических и ассоциативных принципах, при этом после получения очередного пакета сенсорной информации, очищенной, агрегированной и интегрированной предыдущими блоками, задействуются ассоциативные и иерархические связи для формирования на основе текущего среза целостного описания окружающей среды модели ситуации, в которой находится

ИИ-агент, что также включает в себя и контекст. Именно на этом шаге у ИИ-агента формируется понимание того, что происходит.

6. Сформированная модель ситуации рассматривается проактивной подсистемой управления, в которой создаётся набор управляющих воздействий, который отправляется как в блок трансляции команд в код низкого уровня, так и в реактивную подсистему управления для формирования нового рефлекторного контура.

7. Через актуаторы команды, транслированные в код низкого уровня, осуществляют воздействие на окружающую среду, и цикл повторяется с самого начала.

Особенностью этой архитектуры является то, что описанный цикл взаимодействия ИИ-агента со средой, в которой он функционирует, осуществляется непрерывно и, более того, не дожидаясь, пока завершится предыдущая итерация этого цикла. То есть ИИ-агент постоянно воспринимает из среды мультисенсорную информацию, которую подвергает такого рода обработке с формированием у себя разветвлённой ассоциативной памяти, которая в конечном итоге позволит ему понимать происходящее, в том числе и учитывать контекст.

Как видно, описанная схема функционирования понимающего ИИ-агента довольно абстрактна и отвязана от конкретного типа самого ИИ-агента, типа среды и проблемной области. Это значит, что схема является шаблоном, по которому можно строить понимающих ИИ-агентов, конкретизируя их для заданной среды и решаемой задачи в ней. Если возвратиться от абстрактного шаблона к китайской комнате, то описанная схема позволит ИИ-агенту понимать естественный язык в том смысле, как это делает человек, поскольку с самого начала его функционирования он будет нарабатывать свой личный опыт, накапливая ассоциативную память.

Ассоциативная память искусственного когнитивного агента в этом случае будет представлять очень сильно связанную семантическую сеть [Žáček, Telnarová, 2019], активация узлов в которой будет представлять собой «понимание текущей ситуации». Множество всех активированных в момент времени узлов семантической сети является актом когниции, мыслью когнитивного агента. Переход от одного множества активированных узлов к другому в этом случае будет являться «потоком мыслей».

Фактически представленные принципы и шаблонная схема проектирования и реализации понимающего ИИ-агента открывают путь к сильному ИИ или искусственному интеллекту общего уровня (AGI). Здесь важно заострить внимание на том, что в этом случае вполне может так получиться, что созданный ИИ-агент с возможностями мультисенсорной интеграции и ассоциативной памятью сможет

получить внутренние феноменальные состояния [Душкин, 2020]. Однако это остаётся вопросом дальнейших исследований.

Заключение

В настоящей работе показано, что мысленный эксперимент Дж. Сёрла «китайская комната» должен быть пересмотрен и не применяться к современным подходам к построению искусственных когнитивных агентов, которые проектируются и реализуются в рамках гибридной парадигмы искусственного интеллекта. Термин «понимание» не был чётко определён автором мысленного эксперимента, а потому вопрос о том, понимает ли ИИ-агент смысл обращённых к нему фраз на естественном языке, лишён смысла.

Тем не менее, сделана попытка определить феноменологию понимания естественными интеллектуальными агентами с дальнейшим переносом этого определения на агентов искусственной природы. Представлена гибридная архитектура ИИ-агента, основанная на аналогии с верхнеуровневым абстрактным рассмотрением процессов, происходящих в нервной системе высших животных. ИИ-агенты, которые будут спроектированы и реализованы в рамках этой архитектуры, могут стать предтечами к сильному ИИ.

Вместе с тем за рамками настоящей работы остались пока нерешённые вопросы, которые требуют дальнейших исследований. Должны быть рассмотрены принципы организации иерархической ассоциативной памяти для реализации понимания у ИИ-агентов. Также необходимо тщательно проанализировать сходства и отличия базовых компонентов понимания и принятия решений у птиц, в головном мозге которых отсутствует кора, и наземных животных, обладающих корой различных типов и архитектоники. Более того, в область внимания исследователей должны попасть роевые интеллектуальные системы естественной природы – муравьиные кучи, рои пчёл, стаи птиц и т. д.

Список литературы / References

Душкин Р. В. Искусственный интеллект. М.: ДМК-Пресс, 2019.

Dushkin R. V. *Iskusstvennyi Intellekt.* Moscow, DMK-Press, 2019. (in Russ.)

Душкин Р. В. К вопросу о распознавании и дифференциации философского зомби // Философская мысль. 2020. № 1. С. 52–66.

Dushkin R. V. K voprosu o raspoznovanii i differentsiatsii abkjcjacrjuj zombi. *Filosofskaya Mysl*, 2020, no. 1, p. 52–66. (in Russ.)

Душкин Р. В., Андронов М. Г. Гибридная схема построения искусственных интеллектуальных систем // Кибернетика и программирование. 2019. № 4. С. 51–58.

Dushkin R. V., Andronov M. G. Gibridnaya shema postroeniya iskusstvennykh intellektualnykh system. *Kibernetika i Programmirovanie*, 2019, no. 4, p. 51–58. (in Russ.)

Шмидт Р., Тевс Г. Физиология человека: В 3 т. М.: Мир, 1996. Т. 1.

Schmidt R., Thews G. Human Physiology. Transl. into Russian. Moscow, Mir, 1996. (in Russ.)

Шумский С. А. Машинный интеллект. Очерки по теории машинного обучения и искусственного интеллекта. М.: РИОР, 2020.

Shumski S. A. Mashinnyi intellect. Ocherki po teorii mashinnogo obucheniya i iskustvennogo intellekta. Moscow, RIOR, 2020. (in Russ.)

Ashby W. Design for a Brain. The origin of Adaptive Behaviour. New York, Wiley & Sons, 1960.

Bereiter C. Education and Mind in the Knowledge Age. Lawrence Erlbaum Associates, 2009.

Cannon W. Physiological Regulation of Normal States: Some Tentative Postulates Concerning Biological Homeostatics. Paris, Les Éditions Médicales, 1926.

Chang Y. Reorganization and plastic changes of the human brain associated with skill learning and expertise. *Frontiers in Human Neuroscience*, 2014, no. 8, p. 35. DOI 10.3389/fnhum.2014.00035.

Dennett D. Consciousness Explained. Penguin Books, 1991.

Jones E. The Thalamus. Springer, 1985.

LeDoux J. How does the non-conscious become conscious? *Current Biology,* 2020, vol. 30, no. 5, p. 196–199.

Melchitzky D., Lewis D. Functional Neuroanatomy. In: Sadock B., Sadock V., Ruiz P. (eds.). Kaplan and Sadock's Comprehensive Textbook of Psychiatry. Philadelphia, PA, Lippincott Williams & Wilkins, 2009, p. 5–42.

Searle J. Minds, brains, and programs. *Behavioral and Brain Sciences*, 1980, vol. 3, no. 3, p. 417–424.

Searle J. Chinese Room Argument. In: Keil F., Wilson R. (eds.). The MIT Encyclopedia of the Cognitive Sciences. MIT Press, 2001, p. 115–116.

Stout D., Khreisheh N. Skill Learning and Human Brain Evolution: An Experimental Approach. *Cambridge Archaeological Journal,* 2015, vol. 25, no. 4, p. 867–875.

Turing A. Computing Machinery and Intelligence. *Mind,* 1950, vol. 59, p. 433–460.

Žáček M., Telnarová Z. Language networks and semantic networks. In: Trník A., Medved I. (eds.). Proceedings of Central European Symposium on Thermophysics 2019 (Banska Bystrica, Slovakia, October 16–18, 2019). AIP Publishing, 2019, p. 264–269.

Материал поступил в редколлегию
Received
28.01.2020

Сведения об авторе / Information about the Author

Душкин Роман Викторович
директор по науке и технологиям, Агентство искусственного интеллекта, ООО «Дикрипто» (Москва, Россия)

Roman V. Dushkin
Chief science and technology officer, Artificial Intelligence Agency, Deecrypto LLC (Moscow, Russian Federation)

roman.dushkin@gmail.com