УДК 165.731.3

КИТАЙСКАЯ КОМНАТА ДЖ. Р. СЕРЛЯ В КОНТЕКСТЕ ПРОБЛЕМ ФИЛОСОФИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

С.Ю. Нечаев

Саратовский государственный университет
E-mail: donotdespair@yandex.ru

Статья посвящена ключевому моменту в философии искусственного интеллекта – постановке проблемы соотношения синтаксиса и семантики. Анализируется мысленный эксперимент Дж. Р. Серля – Китайская комната, который опровергает всякую возможность создания сильного искусственного интеллекта. Прослеживается история этого вопроса, начиная с теста Тьюринга и заканчивая его интерпретацией С. Харнадом. Высказываются идеи по устранению аргумента Дж. Р. Серля.
Ключевые слова: философия искусственного интеллекта, тест Тьюринга, сильный искусственный интеллект, Дж. Р. Серль, Китайская комната, квалиа, абдукция.

J.R. Searle's Chinese Room in the Problem Context of the Philosophy of AI

S.Y. Nechaev

This article is related to the key moment of philosophy of AI by statement of a correlation problem between syntax and semantics. Author analyzes J.R. Searle's mental experiment called Chinese Room, which denies any opportunity of the "strong AI" creation. The history of this question is retraced, since Turing test till its S. Harnad interpretation. The author suggests express ideas on elimination of J.R. Searle's argument.
Key words: philosophy of AI, Turing test, strong AI, J.R. Searle, Chinese Room, qualia, abduction.

В середине XVIII в. французский философ-материалист Ж.О. де Ламетри анонимно опубликовал свое главное сочинение «Человек-машина», принесшее ему широкую известность и впервые в философии затронувшее проблему соотношения человека и безжизненного аппарата, где человек рассматривался как набор пружин и представлял собой нечто подобное часовому механизму. Наивные взгляды де Ламетри не подтвердило дальнейшее развитие науки, однако он сумел поставить новый вопрос гносеологии, область которой вслед за Р. Декартом именуют дуализмом в философии. Придерживаясь крайнего дуализма, де Ламетри утверждал полное отличие разума и тела и абсолютное подобие механической природы человека всему остальному миру, что сыграло большую роль в становлении общего механицизма.

Спустя два столетия философия вернулась к этой идее, но совершенно с противоположной стороны, а именно – может ли машина в каком-то смысле являться подобием человека, и может ли человек искусственно не только воссоздать тело, но и наделить его разумом. История этого вопроса начинается в середине XX в. с публикации ставшей уже классической статьи английского математика А. Тьюринга «Вычислительные машины и разум», в которой на основе высказанного ранее совместно с американским математиком и логиком А. Черчем тезиса о вычислимости некоторой функции предлагался тест для определения интеллектуальности вещи. Такой *тест Тьюринга* должна пройти гипотетическая *машина Тьюринга*, в ходе которого, опираясь на результаты системы вопросов-ответов (или входов-выходов), при невозможности отличия собеседников – человека от компьютера, мы вынуждены будем признать наличие искусственного интеллекта (ИИ) у испытуемой машины.

Критерий Тьюринга выглядит намного обоснованнее с точки зрения точной науки, по сравнению с утверждениями материалистов Нового времени. Декарт определял дуализм познания в качестве простой интроспекции, тогда как Тьюринг развил бихевиористский подход вслед за философскими концепциями второй половины XIX в. На протяжении почти тридцати лет предложенный Тьюрингом эксперимент не вызывал никаких сомнений, более того, он был подкреплен гипотезой американских ученых А. Ньюэлла и Г. Саймона о правомерности существования компьютера, оперирующего физическими символьными системами. Впоследствии он будет назван *сильным искусственным интеллектом* (strong AI). Создавались первые

прототипы «машин Тьюринга», а писатели-фантасты обрели новую пищу для расцветающего литературного жанра киберпанка.

В 1980 году американский философ Дж. Р. Серль публикует небольшую статью «Сознание, мозг и программы», в которой при помощи созданного им мысленного эксперимента доказывает абсолютную невозможность создания сильного искусственного интеллекта, подобного человеческому мозгу[1]. Эксперимент получил название *Китайской комнаты* и вызвал бурную полемику среди ученых разных стран.

В такую комнату Серль поместил человека *A*, абсолютно ничего не понимающего в китайских символах, но знающего английский язык. В его распоряжении находятся три корзины с «данными»: в первой текст (рукопись) на китайском языке; в следующей – китайские символы и правила на английском, позволяющие сопоставить первую корзину со второй (рассказ); в последней – еще один набор правил на английском для сопоставления первых двух корзин с третьей (вопросы). Собеседник *B*, находящийся вне комнаты и свободно разговаривающий по-китайски, посылает вопросы, на которые получает соответствующие ответы из комнаты. Путем простого оперирования символами трех корзин *A* сможет выдать логически верные ответы на вопросы *B* относительно одного из рассказов. Таким образом, *A* пройдет тест на знание языка, легко убедив в этом своего собеседника, хотя и не понимает ни одного знака в наборе китайских иероглифов.

Представим теперь цифровой компьютер *C*, который на основе своих программ моделирует работу *A*. Точно таким же образом он сможет пройти тест, оперируя системой символов, но не понимая их значения. Аргумент Серля, как видим, направлен против теста Тьюринга и исключает любую возможность создания сильного искусственного интеллекта. Достоинств слабого искусственного интеллекта (weak AI), которые заключаются в простом моделировании человеческой деятельности, оказывается, не хватает для сопоставления ментальности биологического мозга с работой компьютерного процессора. Бихевиористский критерий Тьюринга будет недостаточным, и Серль выдвигает как минимум два тезиса против сильного искусственного интеллекта: 1) синтаксис не определяет семантическое содержание символов, точно так же, как компьютер обращается с данными, не придавая им никакого значения; 2) человеческий разум контролируется мозговой деятельностью, и это биологический процесс, тогда как аппаратное обеспечение (hardware) не соотносится таким же образом с программным обеспечением (software) компьютера.

Статья Серля завершалась вопросами из Йельского и Массачусетского университетов, университета Беркли и других, на которые были получены шесть последовательных ответов[2]. Первый вопрос предполагал, что если человек в комнате не знает языка, то сама комната как система знает китайский язык. Ответ Серля (ответ от систем) показывает, что сама система комнаты ничем не отличается от человека в ней. Мы можем представить, что человек выучит наизусть все символьные данные, предоставленные ему, но он также не будет обладать знанием этих данных, так как овладеет синтаксисом, а не семантикой. Следующий вопрос предполагал наличие компьютера внутри робота, руководящего его движениями изнутри. Серль продолжает утверждать (ответ от робота), что повторение простых моторных функций не соответствует каузальной деятельности человеческого мозга и не может быть расценено как интеллектуальная активность. Мы можем поместить человека в робота, подобно человеку в комнате, и он будет руководить роботом, совершенно не осознавая его общей деятельности, а лишь обрабатывая конкретные входные области данных в выходные данные. Третий основной вопрос заключался в полном моделировании компьютером работы головного мозга человека. Ответ Серля вполне прост: биологический мозг – самый сложный орган в природе, его структура и принципы действия до сих пор до конца не изучены, поэтому идея в точности повторить процессы взаимодействия нейронов и синапсов мозга оказывается еще более абсурдной, чем создание самой интеллектуальной вещи.

Кроме того, этот вопрос странен в том смысле, что сторонники сильного искусственного интеллекта изначально пытались понять, как работает сознание (насколько это возможно), и оно не было связано непосредственно с биохимическими процессами головного мозга.

Как видим, аргумент Серля о Китайской комнате выглядит более чем убедительно и выдерживает всю направленную на него критику. Позиция философа подкрепляется современной аналитической философией сознания, в частности, такой ее важной категорией, как *квалиа*. Вслед за идеями, высказанными в книге американского исследователя Х. Дрейфуса «Чего не могут вычислительные машины» (1972), Серль уделяет много внимания биологическому аспекту человека[3]. Он утверждает, что машины не обладают интенциональностью в смысле аналитической и лингвистической философских традиций, то есть их «сознание» не обусловлено субъективной компонентой. Такой компонентой выступает «квалиа» в качестве свойств чувственного опыта. Другими словами, компьютеры не способны к простому проявлению чувств, которые во многом формируют человеческое сознание. Однако робот-андроид ASIMO, представленный в 2005 г. корпорацией «Хонда», уже способен к распознаванию движущихся объектов, различает звуки и узнает до десяти человеческих лиц, умеет пользоваться Интернетом и локальными сетями.

Утверждение Серля о квалиа сознания отрицается американским философом Полом Черчлендом, который в соавторстве с Патрицией Черчленд в 1990 г. продолжил полемику с Серлем[4]. Черчленды предположили, что нервная система человека может соответствовать архитектуре сети параллельных машин, в ответ на утверждение Серля о том, что импульсы в организме распространяются в миллионы раз быстрее моделируемых импульсов какого-либо робота. Такая сеть, по их мнению, не только сможет ускорить процессы передачи данных, но и позволит обрабатывать их одновременно разными путями, что и происходит в отношении нейронов-синапсов головного мозга. Это избавит систему от большего, по сравнению с ошибками локального компьютера, числа ошибок и поможет в создании кратковременной памяти машины, которая станет формировать собственное «квалиа» компьютера[5].

Знаменитый польский писатель и философ С. Лем в своей статье по поводу комнаты Серля попытался раскрыть ее «тайну», представив несколько измененные варианты этого мысленного эксперимента[6]. Что если запрограммированный компьютер и человек собирают головоломку (puzzle) из кусочков различной формы, подходящих друг к другу только в единственно верном случае: будут ли их ментальные действия схожи между собой? Результатом решения такой загадки может быть картина какого-нибудь художника или обычное слово на обороте собранного «паззла», но ни человек, ни компьютер не знают заранее о нем, пока не соберут все части головоломки. Вдвоем они совершают одни и те же логические операции, что приводит к одинаковому интеллектуальному выводу. Лем сомневается и в значимости «китайского языка» в эксперименте Серля. Если представить, что человек в комнате оперирует символами языка, близкого его «родному», то станет очевидно, что вопросы во многом детерминируют ответы. Поэтому, допустим, из трех вариантов ответов всегда можно выделить один наиболее интеллектуально интересный, и он всегда будет самым абстрактным. Следовательно, мы не можем исходить из простого «ответа» на вопрос, когда нужно установить и сам уровень этого ответа. И, наконец, Лем предлагает радикальный вариант эксперимента, а именно – закрыть самого Серля в «китайской комнате», пройдет ли он собственный тест и каковы будут его результаты?

Неужели мы опять вернулись к тому, с чего начали? Поистине «игры разума» творят чудеса и слишком часто заводят в тупик. На данный момент ни одна из известных машин не приблизилась к прохождению теста Тьюринга, и по оценкам специалистов это произойдет не раньше 2020 г. Мало того, как мы видим, тест Тьюринга (weak AI) недостаточен для определения интеллектуальности вещи, и более того, эксперимент Серля

(strong AI) также в этом отношении неоднозначен. Эта проблема философии искусственного интеллекта находит свое продолжение в виде *полного теста Тьюринга*, который в 1990 г. предложил канадский философ венгерского происхождения С. Харнад[7]. В свое время именно в журнале, основанном им в 1978 г., была напечатана знаменитая статья Дж. Серля.

Харнад следует идеям, которые высказали американские ученые Ю. Черниак и Д. МакДермотт в своей книге «Введение в искусственный интеллект» (1985), предложив «робототехнический» критерий для определения интеллектуальности машины. Полный тест Тьюринга предполагает не только символьное моделирование, но и физическую имитацию человека – обладание, прежде всего, зрением, способностью реагировать на звуки и другими органами чувств, которые будут определять моторные способности и сознательные действия такой механической «копии» человеческого организма. Видимо, полнота интеллектуальных тестов этим не исчерпывается, и в дальнейшем они будут только усложняться. В итоге это приведет философию искусственного интеллекта все к тому же критерию интроспекции Декарта и к последнему экзамену для машины в качестве знаменитого утверждения cogito ergo sum. И, кто знает, возможно, биологические андроиды не только приобретут квалиа сознания, но и станут экзистенциально переживать свое существование, испытывая, к примеру, страх смерти, подобно главным героям знаменитого фильма Р. Скотта «Бегущий по лезвию».

Помимо определения интеллектуальности вещи, но не ее создателем, а самой этой вещью, необходимо учитывать и аспект самообучения машины. Теоретически предполагается, что робот может стать более совершенным объектом, чем сам человек, в том числе, и интеллектуально, так же, как и цифровой компьютер опережает в скорости вычислений работу головного мозга. С этической точки зрения это несет огромную угрозу человеческому существованию в принципе и, несомненно, нарушит знаменитые «три закона робототехники» Айзека Азимова.

Статья Серля о Китайской комнате не только поставила новые вопросы философии искусственного интеллекта, но и в некотором плане притормозила развитие этой области аналитической философии сознания. Позиции ученых разделились, по крайней мере, в виде двух направлений: 1) философы продолжили изучать отношения синтаксиса и семантики, сознания и критериев интеллектуальности; 2) инженеры забросили «тесты Тьюринга» и принялись за создание роботов, имитирующих основные функции человеческого организма.

Однако существенным является последний вопрос, направленный на защиту сильного искусственного интеллекта. Представим компьютер, обрабатывающий данные вычислительного или символьного характера, в какой степени результаты его работы мы сможем утвердить интеллектуально значимыми? Это будут те результаты, которые приводят к новому знанию. Вероятно, такую способность невозможно полностью отрицать: машина на основе своих вычислений абдуктивно может производить гипотезы, которые будут приниматься на пробу. Логика научного открытия во многом исходит из гипотетических предположений, и хотя компьютер не будет осознавать собственных процессов, результат его деятельности можно расценивать как интеллектуальный. Примером может служить программируемая игра в шахматы.

Абдуктивное мышление как научный метод, разработанное американским философом Ч. С. Пирсом, качественно дополняет дедуктивно-индуктивное познание, учитывая творческий элемент, и является, пожалуй, единственным общим критерием ментальности человека и компьютера. В остальном цель аналитической философии сознания, и философии искусственного интеллекта в частности, заключается в создании определенной деятельности машины, которая будет иметь уникальные разумные характеристики, отличные от человеческих мыслительных процессов. Произойдет это или нет, покажет будущее.

Примечания

[1] См.: *Серль Дж.Р.* Сознание, мозг и программы / Дж.Р. Серль // Аналитическая философия: Становление и развитие: Антология / Общ. ред. и сост. А.Ф. Грязнов. М., 1998.

[2] См.: *Harnad S*. Searle's Chinese Room Argument / S. Harnad // Encyclopedia of Philosophy. Macmillan Reference. 2006. Vol.2. P.239–242.

[3] См.: Серль Дж.Р. Разум мозга – компьютерная программа? / Дж.Р. Серль // В мире науки. 1990. №3. С.7–13.

[4] См.: *Черчленд П.М., Черчленд П.С.* Может ли машина мыслить? / П.М. Черчленд, П.С. Черчленд // Там же. С.14.

[5] См.: *Kentridge R.W.* Computation, Chaos and Non-Deterministic Symbolic Computation: The Chinese Room Problem Solved? / R.W. Kentridge // Grounding Symbols in the Analog World with Neural Nets. Think 2: special issue on «Connectionism versus Symbolism» / Eds. D.M.W. Powers, P.A. Flach. 1993. P.44–47.

[6] См.: *Лем С.* Тайна китайской комнаты / С. Лем // Молох. М., 2005. С.246–255.

[7] См.: *Harnad S*. What's Wrong and Right About Searle's Chinese Room Argument? / S. Harnad // Essays on Searle's Chinese Room Argument / Eds. J. Preston, M. Bishop. Oxford, 2001.