ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: PRO ET CONTRA

В.А. Ладов, Н.А. Тарабанов

Данная статья посвящена рассмотрению философских проблем, связанных с развитием систем искусственного интеллекта. Авторы обсуждают работы известных философов аналитической традиции, в первую очередь, исследования Дж. Серла и его знаменитый «Аргумент китайской комнаты». Главная тема этих исследований – сравнение рациональных способностей человеческого сознания и искусственного интеллекта.

ARTIFICIAL INTELLIGENCE: PRO ET CONTRA

V.A. Ladov, N.A. Tarabanov

The article presents a research concerning some philosophical problems of artificial intelligence. Authors investigate works of well-known philosophers of analytic tradition, first of all, John Searle's researches and his famous "Chinese room argument". The main theme of these researches is a comparison of intellectual possibilities of human mind and artificial intelligence.

Вопрос о том, как устроен человеческий мозг и в каком отношении к нему находятся сознание и его ментальные состояния, всегда был и, пожалуй, будет оставаться центральным для целого комплекса наук о познании – когнитивистики. В когнитивистике особое место занимают проблемы, связанные с моделированием тех процессов, которые происходят в мозге человека (прежде всего) как разумного существа, способного мыслить и понимать. Исследования, проводящиеся в данной области когнитивистики, являются неотъемлемой частью такого научного направления как искусственный интеллект (ИИ)[1].

Хотя направление ИИ во многом посвящено разработке машин, которые действуют так, как если были бы «разумны», большинство последних конструируется без всякого намерения подражать когнитивным процессам человека. Однако есть и те, кто занят разработкой «разумных» машин, моделирующих человеческое мышление. При этом не только считается, что машины способны точно копировать человеческое познание, но и утверждается, что наиболее сложные интеллектуальные процессы могут выполняться

[1] Artifical Intelligence. Как известно, такое название было дано Джоном МакКарти - одним из первых специалистов в области компьютерного моделирования [2].

только машинами. Это надо понимать так, что компьютеры должны непосредственно участвовать в повседневном принятии решений людьми [1].

С другой стороны, находятся те, кто полагает ИИ интеллектуально извращенным понятием и считает, что люди, верящие в мыслящие машины, - это материалистические идолопоклонники. Они полагают, что человеческое мышление - это чисто человеческий процесс; наверно, его можно частично синтезировать в машине, но дублировать с помощью компьютерных программ его не удастся никогда [2].

В итоге все многообразие затрагиваемых (но не рассматриваемых во всей своей полноте) проблем зачастую тривиальным образом сводится к одному, предполагающему однозначный ответ, вопросу: «может ли вычислительная машина (цифровой компьютер, искусственным образом смоделированный интеллект и т.д.) мыслить?». Тривиальность такой постановки вопроса нередко усугубляется еще и неясностью в определении довольно часто употребляемых понятий.

Приходится согласиться со следующим замечанием. При установлении некоторого рода параллелизма в работе мозга и компьютера используется весьма много неявных посылок относительно того, как понимать такие термины как «программа» и «материальная часть» (software and hardware) в применении к работе мозга или такие термины как «мышление» или «сознание» в применении к работе компьютера. Концептуальная ясность в таких случаях является чрезвычайно важной, поскольку в противном случае можно погрязнуть в бесконечных спорах по поводу многих, быть может, неправильно поставленных вопросов [3].

Однако здесь не будут особым образом проблематизироваться всякого рода возможные и имеющиеся (в дискуссиях на данную тему) эквивокации, неизбежные с точки зрения специфики той философской позиции, которой неявно придерживается тот или иной исследователь. В данной статье главной целью ставится рассмотрение одной из центральных проблем ИИ (проблемы возможности моделирования человеческого мышления) с критической позиции американского философа Джона Серла (р.1932). Актуальность такого рассмотрения обосновывается тем, что в современной литературе выдвигаемые Серлом аргументы против сторонников ИИ становятся ключевыми и наиболее обсуждаемыми как для его приверженцев, так и для его оппонентов.

Прежде всего, Серл полагает, что необходимо отличать «сильный» ИИ от «слабого» ИИ. Предметом критики у Серла выступает «сильный» ИИ, согласно которому «компьютер — это не просто инструмент в исследовании сознания; компьютер, запрограммированный подходящим образом, на самом деле и есть некое *сознание* в том смысле, что можно буквально сказать, что при наличии подходящих программ

компьютеры *понимают,* а также обладают другими когнитивными состояниями» [4. 377]. Согласно же «слабому» ИИ, компьютеры лишь имитируют мысль, их мнимое понимание не является действительным пониманием, мнимое вычисление – как будто вычисление, и т. д.; тем не менее, компьютерная имитация полезна для изучения сознания (как, например, для изучения погоды и др.).

В противовес «сильному» ИИ Серл предлагает нам провести мысленный эксперимент. Представим себе англичанина, совершенно не знакомого с китайским языком. Его заперли в комнате и дали ему какой-то массивный текст на китайском языке. Затем он получает вторую китайскую рукопись и набор правил на английском языке, чтобы сопоставить первую рукопись со второй. Правила позволяют соотнести один набор формальных символов с другим набором формальных символов; «формальный» (или «синтаксический») означает, что наш англичанин способен полностью распознавать определенные символы по их форме. Наконец, этому англичанину предоставляют третью рукопись, опять же состоящую из китайских символов, и дополнительные инструкции на английском языке, которые позволяют соотносить элементы этой третьей рукописи с элементами первых двух и учат его (англичанина), таким образом, выдавать определенные китайские символы с определенными формами в ответ.

Люди, которые дают эти символы, называют первый текст «рукописью», второй – «рассказом», а третий – «вопросами». Символы, которые выдает англичанин, - «ответами на эти вопросы»; набор правил на английском – «программой». Впоследствии этот англичанин настолько преуспел в следовании этим инструкциям, что с точки зрения кого-то, кто находится за пределами комнаты, его ответы абсолютно неотличимы от ответов носителей китайского языка. Взглянув на эти ответы, никто не осмелился бы сказать, что он ни слова не говорит по-китайски. В действительности же истинно обратное: наш англичанин не понимает (и не знает) китайского. Давая ответы при помощи правил манипулирования неинтерпретированными символами (речь идет о китайском языке), он поступает как компьютер. А именно, как компьютер, работающий по программе «Script Applier Mechanism» (SAM), созданной Шэнком (Schank) и Абельсоном (Abelson) в 1977г. [5] и способной понимать рассказы, - программе, к которой обращается Серл в качестве примера.

Тогда, вопреки «сильному» ИИ, в независимости от того, как ведет себя якобы понимающий компьютер и какая программа заставляет его вести себя этим способом, поскольку те символы, которые он обрабатывает, являются для него бессмысленными (лишенными семантики), то на самом деле он не понимает. В действительности он не мыслит. Его внутренние состояния и процессы, будучи чисто синтаксическими, лишены

семантики (значения); таким образом, в действительности у него нет никаких интенциональных состояний.

Изложив пример и сделав вышеупомянутый вывод, Серл рассматривает несколько ответов, высказанных, когда он имел случай представить этот пример ряду специалистов в области ИИ. Серл предлагает свои возражения на эти ответы.

Ответ от систем (The Systems Reply) допускает, что запертый в комнате человек не понимает рассказа, однако он – только часть целой системы, которая данный рассказ действительно понимает. Серл предлагает позволить этому человеку интериоризировать все элементы системы, путем запоминания правил, производя поиск и другие операции в уме. «И все-таки, - утверждает Серл, - из китайского он ничего не понимает, и… система здесь ни при чем, потому что в системе нет ничего, чего бы не было в нем. Если он не понимает, тогда нет способа, которым могла бы понимать система, потому что система – это только его часть» [6. 420].

Ответ от робота (The Robot Reply) – пользующийся поддержкой современных каузативных теорий значения – предлагает разделять иероглифы и ту действительность, которую, как предполагается, они представляют. Чтобы продвинуть символьную манипуляцию до подлинного понимания, она должна основываться на внешнем мире, через причинные связи человека с теми вещами, к которым иероглифы, в качестве символов, применяются. Если мы помещаем внутри робота компьютер, с тем чтобы использовать робота таким способом, что он делает нечто очень похожее на восприятие, ходьбу, перемещение, тогда этот робот имел бы подлинное понимание и прочие ментальные состояния. Выступая против данного ответа, Серл предлагает провести тот же самый (только немного модифицированный) эксперимент. Поместите внутри робота китайскую комнату и представьте, что некоторые из китайских символов появляются из телевизионной камеры, встроенной в робота, и что другие китайские символы предназначены для того, чтобы служить двигателем внутри робота, двигать его ноги и руки. Однако Серл настаивает на том, что в этом случае мы не понимаем ничего, за исключением правил манипуляции с символами. Он поясняет, что в этом случае у нас нет никаких ментальных состояний соответствующего (интенционального) типа. Все, что мы делаем – это следуем формальным инструкциям по манипулированию формальными символами.

Ответ от имитатора мозга (The Brain Simulator Reply) предлагает нам представить, что компьютером программа (или человек в комнате) не репрезентирует имеющуюся у нас информацию о мире. Она лишь имитирует фактическую последовательность нейронных процессов в синапсах человеческого мозга, когда он понимает китайские

рассказы и дает ответы на них. Тогда мы вынуждены согласиться с тем, что такой механизм понимает рассказы; в противном случае мы вынуждены отрицать, что китайцы понимают эти рассказы, поскольку на уровне синапсов нет никакого различия между компьютерной программой и мозгом китайца. Однако Серл продолжает настаивать на том, что этого все еще недостаточно для продуцирования понимания. Допустим, что вместо перемещения символов, человек оперирует сложным набором водопроводных труб с соединяющими их клапанами. Определяя некоторые китайские символы как вход, программа говорит человеку, какие клапаны он должен выключить или включить. Каждая водопроводная связь соответствует определенному синапсу в китайском мозге, и система в целом установлена так, чтобы после правильного включения всех клапанов, китайский ответ появлялся на конечном выходе ряда труб. Очевидно, что этот человек совсем не понимает китайского языка, и не монтирует никаких водопроводных труб. Проблема с имитатором мозга, как ее определяет Серл, в том, что она имитирует только формальную структуру последовательности нейронных процессов: недостаточность этой формальной структуры для продуцирования значения и ментальных состояний демонстрируется на примере с водопроводными трубами [6. 421].

Комбинированный ответ (The Combination Reply) допускает все выше сказанное: встроенный в робота компьютер, который управляет имитирующей мозг программой и все это рассматривается как единая система. Конечно, теперь мы должны были бы приписать системе интенциональность. В действительности, замечает Серл, как ни один из этих ответов, взятых по отдельности, не имеет никакого шанса опровергнуть полученный им в ходе мысленного эксперимента результат, так и рассматриваемые вместе они не способны сделать это. Данный ответ основывается лишь на том предположении, что если робот выглядит и ведет себя достаточно похожим на нас образом, тогда мы вправе допустить, пока не доказано обратное, что он обладает ментальными состояниями, которые обуславливают и выражают его поведение. Однако если бы мы знали, как объяснить его поведение безотносительно к высказанным предположениям, мы бы не приписывали ему интенциональность. Тем более мы не делали бы этого в том случае, когда знали, что он имеет некую формальную программу [6. 421].

Ответ от других сознаний (The Other Minds Reply) напоминает нам, что, как известно, другие люди понимают китайцев или еще кого-либо, исходя из их поведения. Следовательно, если компьютер также способен пройти испытания на адекватность поведения, как и человек, то в этом случае познание должно приписываться в равной степени им обоим. Серл отвечает, что здесь упускается следующий момент: вопрос не в

том, откуда мне известно, что люди обладают когнитивными состояниями, а в том, *что* именно я им приписываю, когда приписываю когнитивные состояния. Суть аргумента состоит в следующем: вычислительные процессы и их результат могут существовать и без какого-либо когнитивного состояния [6. 420-421].

Ответ от большинства (The Many Mansions Reply) предполагает, что даже если компьютерная программа не способна продуцировать интенциональность и другие когнитивные состояния, то со временем будут изобретены другие средства, при помощи которых это станет возможным. Таким образом, наличие у компьютеров интенциональности и других ментальных состояний – это лишь вопрос времени. Здесь, по мнению Серла, также упускается из виду следующий важный момент. Ответ от большинства упрощает проект «сильного» ИИ, переопределяя его как все, что искусственно производит и объясняет познание, тем самым, отказываясь от первоначального заявления, сделанного от имени искусственного интеллекта, согласно которому умственные процессы есть вычислительные процессы через формально определенные элементы. Если ИИ не отождествляется с этим «достаточно точно определенным тезисом, - пишет Серл, - то мои возражения далее безотносительны, потому что для них более не существует гипотезы, проверяющей их допустимость» [6. 422].

Назвать эксперимент с китайской комнатой спорным было бы преуменьшением. Начиная с возражений, опубликованных с момента первой презентации данного эксперимента Серлом (1980), мнения решительно разделились, не только относительно того, убедителен ли этот аргумент. Среди тех, кот признает его неубедительность, есть и те, кто задается вопросом, почему он неубедителен; а среди тех, кто думает иначе, также есть и те, кто задается вопросом, почему это не так.

Основные возражения и ответы на представленный Серлом аргумент китайской комнаты, помимо вышеназванных ответов, принимают, главным образом, два направления. Одно направление, принимаемое, например Даниелем Деннеттом [7], осуждает имплицитно выраженный в позиции Серла дуализм. Другое направление отмечает, что символы, обозначающие совершаемые англичанином процессы, есть не бессмысленные шифры, а являются китайскими надписями. Следовательно, они очень даже осмысленны и могут быть подвергнуты определенным операциям, обработке, в независимости от того, знает об этом факте англичанин или же нет.

Отвечая на эту вторую разновидность предъявляемых ему возражений, Серл настаивает на том, что спорным моментом в данном случае выступает *внутренняя интенциональность* (intrinsic intentionality), в противоположность *производной*

интенциональности (derived intentionality) надписей и других лингвистических знаков. Все значения вычислений, производимых англичанином, могут быть определены посредством тех значений китайских символов, которые он обрабатывает. При этом символы не будут присущи данному процессу или системе в целом, но будут свойственны «родственному наблюдателю» ("observer relative"), существуя только в сознании наблюдателей, подобных урожденным китайцам, находящимся вне комнаты. «Приписывание интенциональности родственному наблюдателю всегда зависит от внутренней интециональности этих наблюдателей» [8. 451-452]. Суть данного эксперимента, согласно изложению Серла, тогда, в следующем: «реализация программы не может послужить основанием интенциональности, ибо возможно такое, что какой-либо агент [например, наш англичанин], реализуя программу, все еще не имеет должной интенциональности» [8. 450-451]. Хотя Серл отождествляет *внутреннюю* интенциональность с интенциональностью сознания, тем не менее, он способен противостоять обвинениям Деннета в инсинуациях дуализма. Учитывая тот немаловажный момент, что интенциональность сознания непосредственным образом связана с наличием у нас определенных ментальных состояний, Серл утверждает, что отстаивание «первоначальной точки зрения» здесь обоснованно. Ибо «онтология сознания есть онтология в первой инстанции»: «сознание состоит из qualia [субъективных опытов сознания]…которые лежат в его основании» [9. 20]. Этот тезис онтологической субъективности, как его называет Серл в своих более поздних работах, *не* есть (что он особо подчеркивает) какая-то дуалистическая инсинуация дискредитированного «Картезианского аппарата» [9. xii], как утверждают его критики. Это лишь подтверждает то, что на долгое время было своевольно отклонено бихевиористскими воззрениями и их функционалистскими сторонниками. Это вполне здравое отождествление мысли с сознанием, как утверждает Серл, без труда может быть совместимо с радикальным физикализмом, когда мы представляем сознание как то, что обусловлено основными мозговыми процессами и в них реализовано. Исходя из этого, по мнению Серла, очевидно, что отождествление мысли с сознанием не есть разновидность дуализма; более точно это можно определить как монистический интеракционизм [8. 455-456] или «биологический натурализм» [9. 1].

Таким образом, Серл приходит к формулировке «вывода из аксиом», которая резюмирует главные результаты его мысленного эксперимента. Этот вывод исходит из следующих трех аксиом:

Аксиома 1. *Компьютерные программы - это формальные (синтаксические) объекты.*

Аксиома 2. *Человеческое сознание оперирует смысловым содержанием (семантикой).*

Аксиома 3. *Синтаксис сам по себе не составляет семантику и его недостаточно для существования семантики.*

Отсюда делается следующий очевидный вывод:

Заключение 1. *Программы не являются сущностью сознания, и они недостаточны для его наличия.*

Далее Серл добавляет четвертую аксиому:

Аксиома 4. *Мозг порождает сознание.*

из которой, как предполагается, мы «немедленно получаем тривиальное» заключение:
Заключение 2. *Любая другая система, способная порождать сознание, должна обладать казуальными свойствами, (по крайней мере) эквивалентными соответствующим свойствам мозга.*
И далее:

Заключение 3. *Любой артефакт, порождающий ментальные явления, любой искусственный мозг должен иметь способность воспроизводить специфические казуальные свойства мозга, и наличия этих свойств невозможно добиться только за счет выполнения формальной программы*.

Заключение 4. *Тот способ, посредством которого человеческий мозг на самом деле порождает ментальные явления, не может сводиться лишь к выполнению компьютерной программы* [10. 27-29]

Совершенно ясно, что эксперимент с китайской комнатой приводит к такому выводу посредством «признания третьей аксиомы» [11. 34]. Это позволяет одним из исследователей ИИ сформулировать возможный коннекционистский ответ мысленному эксперименту Серла. Утверждается, что мнимый результат этого мысленного эксперимента достоверен лишь постольку, поскольку в нем самом отсутствует его нейрофизиологическое правдоподобие. Данный ответ призывает нас представить более адекватное мозгу *коннекционистское* его построение. Вообразите Серла-в-комнате (Searle-in-the-room), в этом случае он только один из очень многих агентов, работающих параллельно, при этом каждый вносит свою лепту в общий процесс обработки данных

(подобно множеству нейронов мозга). Поскольку Серл-в-комнате, в этом пересмотренном сценарии, выполняет лишь очень малую часть большой вычислительной работы по производству осмысленных китайских реакций в ответ на информацию, поступающую из китайского входного устройства, то естественно, что сам он не постигает процесс целиком; так что едва ли мы должны ожидать, что он схватит или осознает значения тех связей, в которые он включен второстепенным образом, в процессе обработки данных.

Серл возражает, что коннекционистский ответ – включающий элементы как ответа от систем, так и ответа от имитатора мозга – может быть, подобно этим предшествующим ответам, решительно опровергнут путем соответствующей корректировки сценария мысленного эксперимента. Серл предлагает представить китайскую гимназию с множеством людей, разговаривающих только на английском языке, работающих параллельно и производящих на выходе информацию, неотличимую от той, которую производят настоящие китайцы: каждый следует своему собственному (более ограниченному) набору инструкций на английском языке. Однако очевидно, настаивает Серл, что ни один из этих индивидов не понимает; также как взятые вместе они на это не способны. Интуитивно представляется совершенно очевидным, поясняет Серл, что никто и ничто в пересмотренном эксперименте с китайской гимназией не понимает китайских слов ни индивидуально, ни совместно. В китайской гимназии как совместно, так и индивидуально не делается ничего, кроме бессмысленных синтаксических манипуляций, из которых не могла бы с очевидностью возникнуть интенциональность, следовательно, и имеющая значение мысль тоже.

Таким образом, поскольку программа определена в терминах вычислительных операций на чисто формально определенных элементах, то эти операции сами по себе никак не связаны с пониманием.

ЛИТЕРАТУРА

1. Newell A. Intellectual Issues in the History of Artificial Intelligence // The Study of Information: Interdisciplinary Messages, Machiup F., Mansfield U. (eds.), New York: Wiley, 1983, pp. 196–227.
2. Солсо Р. Мышление и интеллект – естественный и искусственный // http://www.humans.ru/humans/87672
3. Старикова И. В. Роль мысленных экспериментов в понимании природы сознания // http://www.philosophy.nsc.ru/journals/philscience/6_99/05_starikova.htm

4. Серл Д. Мозг, сознание и программы // Аналитическая философия: Становление и развитие (антология). – М.: «Дом интеллектуальной книги», «Прогресс-Традиция», 1998. – С. 376-400.
5. Schank R. C., Abelson R. P. Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Eribaum, 1977.
6. Searle, John R. Minds, Brains, and Programs // Behavioral and Brain Sciences, 1980, № 3, pp. 417-424.
7. Dennett, Daniel. The Milk of Human Intentionality // Behavioral and Brain Sciences, 1980, №3, pp. 429-430.
8. Searle, John. Intrinsic Intentionality // Behavioral and Brain Sciences, 1980, № 3, pp. 450-456.
9. Searle, John. The Rediscovery of the Mind, Cambridge, 1992, MA: MIT Press.
10. Searle, J. R. Is the Brain's Mind a Computer Program? // Scientific American, 1990, №262, pp. 26-31. (Русский перевод статьи в журнале «В мире науки», 1990, № 3, с. 7–13 или на http://az13.mail333.com/mat/in1.htm).
11. Churchland, Paul, and Patricia Smith Churchland. Could a Machine Think? // Scientific American, 1990, №262, pp. 32-39. (Русский перевод статьи в журнале «В мире науки»,1990,№ 3, с.14–21 или на http://az13.mail333.com/mat/in1.htm)