«Искусственный интеллект» как аргумент в споре о сознании

И.Ф. МИХАЙЛОВ

В статье рассматриваются история и теория «искусственного интеллекта» в контексте философских дискуссий о сознании и мышлении на Западе и в России. Метод мысленного эксперимента позволяет автору сформулировать и защитить концепцию коммуникативного функционализма в философской рефлексии сознания и его функций.

Ключевые слова: *искусственный интеллект, философия сознания, мозг, коммуникация.*

В настоящей статье не рассматривается актуальное состояние исследований «искусственного интеллекта» (ИИ), не содержатся в ней и какие-либо философские выводы, навеянные этими исследованиями. Я лишь использую гипотезу принципиальной возможности ИИ для обсуждения известных проблем философии сознания.

Постараюсь наметить ответы на два вопроса:

1. Может ли гипотеза ИИ внести свой вклад в традиционные философские дискуссии вокруг сознания и мышления?

2. Может ли философия с помощью собственного неэмпирического

инструментария предложить ответ на вопрос о возможности искусственного моделирования человеческого разума?

Термин «искусственный интеллект» был введен Джоном Маккарти в 1956 г. для обозначения науки и инженерных практик создания «умных машин». В настоящее время среди множества диверсифицировавшихся направлений можно выделить логическое программирование (см. интересную работу Поспелова[1]), нейронные сети[2] и мультагентные системы[3].

Вокруг теста Тьюринга

В 1950 г. Алан Тьюринг[4] сформулировал принцип идентификации машинного интеллекта, который вошел в историю как тест Тьюринга (ТТ). В эпоху Интернета читатель без труда найдёт описание условий этого теста с помощью поисковых систем или Википедии.

Тест вызвал и продолжает вызывать критические атаки. Наиболее серьезными аргументами кажутся соображения о логической связи теста и теорем Гёделя о неполноте, а также предполагаемая неспособность концепции, лежащей в основе ТТ, справиться с проблемой qualia. Из теорем Гёделя о неполноте следует, что всегда существует хотя бы одна формула, неразрешимая для машины, но очевидная для человека[5]. Следовательно, аналитически верно, что машина (или исполняемая в ней программа) не может быть моделью человеческого интеллекта, поскольку, согласно формулировке Дж. Вебба, «из высказывания “Я могу найти ограничения в любой машине” несомненно следует, что я не машина»[6].

Второй аргумент рассматривает компоненты сознания, принципиально не доступные для компьютерного моделирования. К ним относятся, в частности, qualia, под которыми подразумеваются «реальные» субъективные ощущения, переживания и эмоции, как они чувствуются их носителем. Согласно определению Н.С. Юлиной, «в

---

[1] *Поспелов Д.А.* Десять «горячих точек» в исследованиях по искусственному интеллекту // Интеллектуальные системы (МГУ). 1996. Т. 1, вып.1–4. С. 47–56.

[2] *Hopfield J.J.* Neural Networks and Physical Systems with Emergent Collective Computational Abilities // Proc. NatL Acad. Sci. USA Biophysics. 1982. Vol. 79. P. 2554–2558.

[3] *Scheutz M., Andronache V.* The Apoc Framework for the Comparison and Evaluation of Agent Architectures // Proc. of Aaai Workshop on Intelligent Agent Architecture. 2004. P. 66–73.

[4] *Turing A.M.* Computing Machinery and Intelligence // Mind. 1950. № 49. P. 433–460.

[5] *Lucas J. R.* Minds, Machines and Gödel // Minds and Machines ; ed. by A.R. Anderson. Englewood Cliffs, NJ: Prentice Hall, 1964. P. 14.

[6] Цит. по: *Карпенко А.С.* Современное состояние исследований в философской логике // Логические исследования, 2010. № 3.

общей форме можно сказать, что квалиа есть то, каким образом что-то выглядит для нас, кажется нам, в каком качестве оно предстает перед нами»[7]. Вопрос о qualia уже довольно долго разделяет философов аналитической традиции на две группы: (условно говоря) «бихевиористов», настаивающих на иррелевантности qualia для теории значения (Л. Витгенштейн[8] и его последователи[9]), с одной стороны, и более традиционно ориентированных «философов сознания» (Дж. Сёрл[10], Дж. Перри[11] и др.), часто именно в qualia усматривающих суть последнего (а не в формальных правилах и операциях с символами). Бихевиористы показывают, что теория значения – а именно она видится содержательной и методологической основой теории сознания – может быть построена без обращения к «психологическим» сущностям. Философы сознания, напротив, настаивают, что именно «внутренняя жизнь» психики (мозга) порождает смыслы, без которых формальные операции с символами не могут рассматриваться в качестве мышления. Очевидно, что Тьюринг скорее склонялся к первой точке зрения.

В качестве концепции, «играющей на стороне» ТТ, можно рассмотреть гипотезу систем физических символов (СФС), сформулированную Ньюелом и Саймоном[12] в 1976 г., в соответствии с которой любая система физических символов (физических объектов, представляющих другие объекты в соответствии с некоторыми правилами именования) обладает необходимыми и достаточными средствами для сознательного действия. Наверное, можно сказать, что позиция Д. Деннета в его споре с Сёрлом[13] в значительной мере основывается на идеологии СФС.

«Сильный ИИ» и «китайская комната»

Сёрл в свою очередь рассматривает концепции, сводимые к СФС, как «сильный [принцип] ИИ»[14], в соответствии с которым любая программа, эффективно имитирующая интеллектуальные действия, и

---

[7] *Юлина Н.С.* Головоломки проблемы сознания. М., 2004.

[8] *Wittgenstein L.* Philosophical Investigations. Oxford : Blackwell, 1968. P. 257 and on.

[9] *Малколм Н.* Состояние сна. М., 1993.

[10] *Searle J.* Minds, Brains, and Programs // Behavioral and Brain Sciences. 1980. № 3. P. 417–457.

[11] *Perry J.* Knowledge, Possibility and Consciousness. L. ; Cambridge : The MIT Press Cambridge, 2001.

[12] *Newell A., Simon H.A.* Computer Science as Empirical Enquiry: Symbols and Search // Communications of the ACM. 1976. № 19(3). P. 113–126.

[13] http://www.scribd.com/doc/45344182/Dennett-vs-Searle.

[14] *Searle J.* Op. cit.

есть собственно интеллект. Он отличается от «слабого [принципа] ИИ», который допускает, что для моделирования ИИ, возможно, понадобится определенный физический субстрат, искусственные нейронные сети, возможно, даже квантово-вероятностные эффекты, но все же оно в принципе возможно.

Против «сильного ИИ» Сёрл выдвигает аргумент «китайской комнаты»[15] – мысленный экперимент, найти описание которого читателю также не составит труда. Согласно Сёрлу, вся «китайская комната» как система способна пройти ТТ, хотя ни сидящий внутри человек, ни система в целом не понимают китайского. Метафора очевидна: компьютерная программа, умеющая формально оперировать с символами, понятными человеку, не содержит «интеллекта».

Аргумент Сёрла связан с его общей концепцией сознания как системы интенциональных состояний (знания, сомнения, желания и т.п.)[16], суть которых состоит в их направленности на предмет, но они аналогичны qualia в том смысле, что необходимо связаны с субстратом мозга и как таковые воспринимаются исключительно «в первом лице»[17]. Мораль аргумента «китайской комнаты» очевидна: машина, основанная на СФС, не может иметь интенционального состояния, соответствующего пониманию.

По ходу изложения критической концепции Сёрла невозможно не отметить, что она основана на некоторых некритических интуициях, на что указывал, кстати, и соавтор Деннета Д. Хофштадтер[18]. В частности, как мне кажется, Сёрл, подобно карточному фокуснику, помещает в центр ситуации человека, что называется, для отвода глаз. Ведь именно человеку мы интуитивно приписываем свойство понимания, и если в данном случае у него оно отсутствует, то вроде бы искать его больше негде. Тогда как на самом деле в данной ситуации субъектом понимания (выражаясь традиционным философским языком) является, конечно, *справочник*, а человеку отведена техническая роль информационного транспорта. И в том примитивном виде, как он описан Сёрлом, справочник как раз вряд ли пройдет ТТ – ведь мы помним, что собеседник там должен продемонстрировать способность ориентироваться в модальностях человеческого общения: жаргон, шутка, ложь и т.п. Если же справочник будет составлен достаточно тонко для того, чтобы на серьезные китайские выражения отвечать

[15] *Searle J.* Op. cit.

[16] *Hofstadter D.* Reflections on Searle // The Mind's I; D. Hofstadter and D. Dennett (eds.). N.Y. : Basic Books, 1981. P. 373–382.

[17] Его понимание интенциональности резко расходится с таковым Деннета, который рассматривает ее не онтологически, а как одну из стратегий объяснения функционирования сложных систем (см. об этом: *Юлина Н.С.* Головоломки проблемы сознания. С. 104–108). В рамках концепции Деннета, конечно, никакие аналогии между интенциональностью и qualia невозможны.

[18] *Hofstadter D.* Op. cit.

серьезно, а на шутливые – отшучиваться, причем с возможностью случайного (произвольного?) выбора доступных вариантов ответа, тогда почему бы и нет? Все дело в степени сложности программирования.

Субстанциализм и функционализм

Под субстанциализмом в рамках и для целей данной статьи предлагается понимать широкий круг концепций в философии сознания – от традиционного материализма, идеализма, субстанциального дуализма (картезианства) до более современных физикалистского редукционизма, теории «тождества типов» и др., общим для которых является признание необходимой зависимости сознания и его свойств от его же субстрата-носителя, как бы последний ни понимался, равно как и неразрывной онтологической связи между ними.

Напротив, термином «функционализм» предлагаю обозначить семейство концепций, пытающихся найти объяснение ментальным событиям при помощи выявления их функциональных зависимостей от ряда других – ментальных и нементальных – событий, отвлекаясь от их возможных онтологических экспликаций. Хороший обзор различных версий функционализма содержится в статье Д.В. Иванова[19]. Любопытно также, что, по мнению Джегуона Кима, «функционализм есть часть более широкого бихевиористского подхода к сознанию и может быть понят как обобщенная и усложненная версия бихевиоризма»[20].

Согласно любопытному pdf-документу, который можно найти в ряде мест в Интернете[21], – он представляет собой email-переписку между Сёрлом, Деннетом и редактором научного журнала, состоявшуюся в 1997 г. и собранную вместе Деннетом, – главный аргумент Сёрла, которым он защищается от обвинений в примитивном субстанциализме, состоит в том, что биохимический субстрат мозга обладает «достаточными каузальными силами» для того, чтобы причинно обусловить сознание. Проще говоря, он необходим и достаточен для сознания. (Деннет настаивает, что в одной из публичных лекций Сёрл употребил метафору «секреции» по отношению к связке «мозг–сознание», сравнив её с отношением молочной железы и молока, чем немало позабавил Деннета и Хофштадтера.) Можно сказать, что позиция Сёрла в основе своей имеет доктрину субстанциализма, а позиция Деннета может быть интерпретирована как функционализм.

[19] *Иванов Д.В.* Функционализм. Метафизика без онтологии // Эпистемология и философия науки. 2010. № 2.

[20] *Kim J.* Philosophy of Mind. Brown University: Westview Press. A Subsidiary of Perseus Books, L.L.C, 1998.

[21] http://www.scribd.com/doc/45344182/Dennett-vs-Searle.

Деннет воспроизводит известную аналогию противников субстанциализма: одно время полет так же считался функцией «биологического субстрата» птицы, и это убеждение ничем не помогало братьям Райт в их конструкторских усилиях, пока они – вполне в духе функционализма – не подошли к вопросу с точки зрения законов аэродинамики и конечной цели строительства летательного аппарата, отказавшись от попыток прямой имитации природы. Контрвозражение Сёрла: он имел в виду не «секрецию» в буквальном смысле слова, а «каузальные силы», которые, по его мнению, могут содержаться не только в биологическом субстрате мозга: любой его заменитель должен обладать эквивалентными «каузальными силами» для производства сознания. Ответа Деннета на это контрвозражение в рукописи не содержится, но мы можем попытаться сделать это за него. Что значит быть причинно зависимым от субстрата? Когда специалисты по ИИ (не философы) пытаются делать то, что делали братья Райт на начальном этапе – а именно, имитировать природу, – они создают так называемые нейронные сети, которые по сути представляют собой те же компьютерные программы, только более сложные. Тогда загадочная причинная зависимость от субстрата, демистифицированная в эксперименте, оказывается все той же функциональной зависимостью от программы, а проблема воспроизводимости из непреодолимого теоретического предела превращается в вопрос технического искусства.

Однако Деннет и Сёрл, будучи противниками в одном контексте, становятся по одну сторону баррикад в дискуссии с М. Беннетом и П. Хекером[22], которые с витгенштейновских позиций критикуют как «философов сознания», так и значительную часть нейрофизиологов за засорение научного языка иллюзорными субстанциалистскими терминами. Они рассматривают это как своего рода неокартезианство, которое на место субстанциального дуализма Р. Декарта ставит структурный дуализм тела – мозга, приписывая ментальные предикаты только мозгу. Беннет и Хекер возражают против этого варианта дуализма в духе витгенштейновской концепции «глубинной грамматики»: когда мы говорим «я знаю» или «он потерял сознание», мы не подразумеваем мой или его мозг в качестве подлежащего. Попробуем заменить личные местоимения в этих выражениях на «мой/его мозг» – получим бессмыслицу. Свой методологический бихевиоризм авторы демонстрируют в следующем рассуждении о знании: «Если животное знает нечто, оно может действовать и отвечать на стимулы, получаемые из среды, так как не могло бы действовать и отвечать в отсутствии этого знания; если же оно действует таким образом, оно обнаруживает знание. Можно сказать, что мозг является транспортным средством этих способностей, но это означает лишь то, что в от-

[22] *Bennett M., Dennett D., Hacker P., Searle J.* Neuroscience and Philosophy. Brain, Mind, and Language. N.Y. : Columbia University Press, 2007.

сутствии соответствующих нейронных структур животное не смогло бы делать то, что может делать при их наличии. Нейронные структуры мозга отличны от способностей животного, а функционирование этих структур отлично от применения животным его способностей. Короче говоря, знающий есть также агент действия, и его знание проявляется (is exhibited) в том, что он делает»[23]. Бихевиористские аргументы, равно как и уместность бихевиористского прочтения витгенштейновских идей, мы обсудим ниже.

Дубровский vs. Ильенков: спор внутри субстанциализма

Начиная с 1960-х гг. в СССР на фоне реабилитации кибернетики и построения первых «ЭВМ» размером с многоэтажный дом возникает своего рода романтический культ НТР и ее результатов, повлекший за собой безудержный оптимизм в отношении перспектив синтетического моделирования основных человеческих способностей.

Конечно, проблему начали разрабатывать математики: А.Н. Колмогоров[24], Д.А. Поспелов[25] и др. Но тогдашние отечественные философы внесли свой вклад в этот спор прежде всего в виде памятного поединка по поводу философской экспликации мышления и вообще «идеального» между Э.В. Ильенковым и Д.И. Дубровским. Уточню, что главным предметом спора были именно те или иные определения философских категорий, а не прагматический вопрос о возможностях синтетического моделирования человеческого интеллекта. Но, учитывая яркую «антикибернетическую» позицию Эвальда Васильевича по этому вопросу, выраженную, в частности, в книге «Об идолах и идеалах»[26], можно предположить, что концепция его оппонента должна была бы внушать больший оптимизм сторонникам ИИ.

Я бы сказал, что по некоторым параметрам этот принадлежащий истории спор все еще не является формально завершенным. Так, в предисловии к переизданию своей книги в 2002 г. (т.е. фактически в наше время) Давид Израилевич пишет, что для него ряд ее положений сохраняет актуальность, за исключением нескольких ритуально-идеологических абзацев, которые он вычистил[27]. Раз так, то и мне,

[23] Ibid. P. 150.

[24] См., например: *Колмогоров А.Н.* Автоматы и жизнь // Кибернетика ожидаемая и кибернетика неожиданная. М., 1968. С. 10–29.

[25] *Поспелов Д.А.* История искусственного интеллекта до середины 80-х годов // Новости искусственного интеллекта. 1994. № 4. С. 74–95.

[26] *Ильенков Э.В.* Об идолах и идеалах. Киев : Час-Крок, 2006.

[27] *Дубровский Д.И.* Проблема идеального. Субъективная реальность. М., 2002. С. 4.

наверное, позволено выступить с некоторыми комментариями к позициям участников.

В упомянутой книге Ильенков пишет: «Дело в том, что мыслящее существо необходимо должно быть подвижным, и не просто подвижным, а и умеющим активно действовать в согласии с формой и расположением всех других тел и существ. Во-вторых, оно должно активно изменять, переделывать окружающую его естественную среду, строя из нее свое “неорганическое тело” – тело цивилизации»[28]. Человеческая рука или манипулятор, аналогичный по степени свободы действий, есть «необходимое условие» мышления. В этом, по мысли автора, и состоит «настоящий материализм» в отличие от видимого материализма объяснений мышления через мозг. «Ибо работающий человеческий мозг сам является продуктом труда, а наличие самого лучшего в анатомо-физиологическом отношении мозга еще вовсе не гарантирует наличия мышления»[29]. То есть мозг, согласно Ильенкову, – необходимое, но не достаточное условие интеллекта.

Критика Дубровским позиций Ильенкова, в частности его более поздней концепции идеального как всеобщих форм представленности одних материальных предметов в других, но тоже внешних по отношению к телу человека[30], строилась вокруг следующих основных соображений. Во-первых, по его мнению, отказывая в идеальности не интеллектуальным, но субъективным явлениям, таким, как мимолетные образы или чувства, Ильенков должен был бы, согласно железной системе понятий советского диамата, объявить их материальными. Во-вторых, Дубровский приводит множественные примеры «представленности» одних явлений в других в живой и неживой природе, которые Ильенков, по его мнению, должен был бы считать идеальными[31].

Сам же Дубровский развивал «информационный подход» к проблемам сознания, мышления, идеального и т.п., который формулируется им в следующих «ясных, как солнце» тезисах: (1) информация есть отражение одних материальных систем в других: (2) она не существует вне и помимо материального субстрата, который является также и ее «кодом» и по отношению к которому (или которым) она инвариантна; (3) информация в отношении своего субстрата может выполнять функцию управления, что понимается автором на основе концепции «информационной причинности»[32] (количество тезисов мною несколько сокращено за счет сжатия смысла). Далее совершается логический переход к интересующей нас теме через положения «всякое явление сознания есть информация» и «всякое явление созна-

[28] *Ильенков Э.В.* Цит. соч. С. 234.

[29] Там же. С. 235.

[30] *Ильенков Э.В.* Проблема идеального // Вопросы философии. 1979. № 6, 7.

[31] *Дубровский Д.И.* Цит. соч. С. 44–54.

[32] Там же. С. 137.

ния есть функция головного мозга»[33]. Таким образом, согласно Дубровскому, сознание относится к мозгу как информация к своему материальному носителю («субстрату»). В представленных тезисах, по-видимому, не содержится ничего принципиально отличного от стандартной диаматовской «теории отражения», суть которой была рассмотрена мною в другой работе[34], за исключением интригующего тезиса (3). Я бы задумался по его поводу: как отражение, причиненное одной материальной системой другой материальной системе, может в дальнейшем причинять еще что-то этой второй материальной системе, которою оно еще и «кодируется»? Не присутствует ли здесь некоторая *магия отражения* или *магия информации*? Чуть далее[35] автор объясняет это «цепью кодовых преобразований», что, на мой взгляд, вряд ли является объяснением, поскольку ничем не обосновывается необходимость – естественная или логическая – таких преобразований. Ну да не это интересует нас в первую очередь.

На первый взгляд, в отношении возможности ИИ позиции диспутантов должны были бы распределиться следующим образом: Ильенков – «против», поскольку «идеальное» приписывается только человеку вкупе с опредмеченными формами его культуры, тогда как машинное моделирование интеллекта имплицитно мыслится им только как имитация деятельности мозга. Дубровский же на том же основании должен был бы быть «за». Но вот что интересно: концептуально ничто не мешает построить компьютерную модель интеллектуального решения задач как мультиагентную систему, которая воспроизводила бы не только необходимость коммуникации сознательных агентов друг с другом, но и эволюционное развитие каждого из них в отдельности и всей системы в целом. Более того, вроде бы такие исследования ведутся на вполне эмпирическом уровне[36], хотя Эвальд Васильевич мог не знать об этом по понятным причинам. В то же время предполагаемый ИИ-оптимизм Дубровского тоже может разбиться о его же концепцию мозга как «собственной» (т.е. привилегированной) кодирующей системы сознания[37]. Ведь если сознание в собственном, «субъективном» выражении, включающем «творческую активность Я» как способность высокоорганизованной нейронной системы к самоуправлению, возможно только «в материале» мозга, то компьютеры должны разделить второстепенную техническую роль дополнительных кодирующих систем вместе с книгами, картинами и прочими артефактами.

---

[33] Там же. С. 138.

[34] *Михайлов И.Ф.* Наследие советского «критического марксизма» в контексте проблемы мышления // Вопросы философии. 2011. № 5. С. 108–118.

[35] *Дубровский Д.И.* Цит. соч. С. 153.

[36] *Scheutz M., Andronache V.* Op. cit. P. 66–73.

[37] *Дубровский Д.И.* Цит. соч. С. 141, 147–148.

Если же эту высокоорганизованную нейронную систему в принципе возможно моделировать в металле, силиконе или в чем-то еще, отличном от серого вещества (а, собственно, почему бы и нет – такие исследования тоже ведутся[38]), то человеческий мозг лишается своего априори особого положения как привилегированной кодирующей системы сознания (этот выбор, как мы увидим далее, подробно обсуждается в споре между Сёрлом и Деннетом). Но тогда падает последний (и по сути единственный) бастион защиты от ильенковского подхода к идеальному: если «собственной кодирующей системой» сознания может быть железка, то почему ею не может быть целостность кодирующих (в смысле Ю.М. Лотмана) систем культуры? А естественный «бортовой компьютер» индивида может быть понят тогда как технический инструмент его ассимиляции в культуре и – да, преобразования ее, только именно как инструмент, а не субъект. (Аналогия: бортовой компьютер автомобиля в чем-то помогает ему ехать, но не едет вместо него.)

Таким образом, с одной стороны, выбор культуры или мозга в качестве кодирующей системы сознания не влечет принятия соответственно пессимистической или оптимистической позиции в отношении ИИ, а с другой стороны, концептуальное исследование возможности искусственного интеллекта также не дает преимуществ ни одной из рассматриваемых позиций.

Сравнительный анализ «измов»

Попытаемся теперь обобщить различные позиции относительно необходимых и достаточных условий интеллекта (мышления, сознания – пусть пока они будут синонимами). В качестве таковых мы можем рассматривать альтернативно: (1) культуру; вслед за Ильенковым (2) естественное или искусственное воплощение вычислительной программы, манипулирующей символами (Деннет, Хофштадтер); (3) биохимию мозга (Сёрл, Дубровский – по крайней мере как его концепция представлена в «Проблеме идеального»[39]). (1) и (3)

[38] *Hopfield J.J.* Op. cit. P. 2554–2558.

[39] В более поздней работе Дубровский в связи с критикой Сёрла объявляет себя сторонником функционализма, однако последний видит «в качественном разграничении отношений функциональных и физических (с учетом необходимой связи первых со вторыми), в отрицании редукции функционального к физическому, в обосновании особого типа каузальности и закономерностей, не сводимых к физическим, что имеет принципиальное значение для исследования самоорганизующихся систем (где главная роль принадлежит расшифровке кодовых, т.е. функциональных, зависимостей). Эта суть выражается принципом воспроизводимости одной и той же функции на различной субстратной основе и принципом инвариантности информации по отношению к физическим свойствам ее носителя» (*Дубровский Д.И.* Сознание, мозг, искусственный интеллект. М., 2007. С. 67). Мне, честно говоря, пока не ясно, как эта установка сочетается с уже приводившимися мыслями Давида Израилевича об уникальности мозга как субстрата сознания (*Дубровский Д.И.* Проблема идеального. С. 141, 147–148).

можно временно представить в качестве разновидностей субстанциализма. Добавим еще одно измерение в виде различия холизма (акцент на социальном целом) и индивидуализма и получаем:

	Субстанциализм	Функционализм
Индивидуализм	(3)	(2)
Холизм	(1)	?

Позицию, обозначенную знаком вопроса, сформулируем в конце статьи.

Возражение против (1) и (3): что значит быть функцией субстрата? Сам субстрат, каким бы он ни был, есть (или, по крайней мере, понимается в рамках современной научной картины мира как) определенная структура. Тогда мышление – это функция социальной или атомно-молекулярной или еще какой-нибудь структуры. И тогда любое утверждение эксклюзивности структуры определенного типа в качестве претендента на роль субстрата ментальных событий очевидно несет печать догматизма – или же оно должно быть солидно эмпирически обосновано, что, конечно же, выходит за пределы компетенций философии.

Позиция (2) Деннета, по-видимому, состоит в том, что мышление есть программа, осуществляемая в мозге, но могущая быть реализована в любом субстрате (уже приведенная выше аналогия с полетом). Однако, согласно Беннету и Хекеру, локализация мышления в мозге приводит (в том числе и значительную часть нейрофизиологов) к своего рода *неокартезианству*, когда роль субстанций выполняют тело и мозг, и при этом ментальные предикаты приписываются только последнему. Однако знает, верит, теряет сознание и падает в обморок именно человек (животное), а не мозг. Да, это всего лишь «внутренняя грамматика» языка, но она задает онтологию. Можно сказать, что в случае с потерей сознания употребление «я» – лишь языковая конвенция: на самом деле речь идет о неполадках в мозге. Но можно ли сказать, что «знать» и «думать» – это тоже предикаты мозга?

Беннет и Хекер интерпретируют Витгенштейна бихевиористски: знание есть его проявление в поведении. Насколько такая интерпретация справедлива, надеюсь, станет ясно из дальнейшего изложения. Для этого необходимо обсудить понятие интенциональности, введенное в аналитическую традицию ученицей Витгенштейна Г.Э.М. Энскомб[40] и развиваемое в дальнейшем Сёрлом[41].

Итак, позиция (3) тесно связана с концепцией интенциональности. Сознание – не столько содержание, сколько направленность на

[40] *Anscombe E.* Intention. Ithaca, NY : Cornell University Press, 1957.

[41] *Searle J.R.* Consciousness and Language. Berkley: University of California, 2002. P. 77–90.

предмет, или, как ее интерпретируют некоторые англоязычные авторы, aboutness. Согласно Сёрлу, сознание производится биологическим субстратом мозга. Следовательно, главная работа мозга – производство интенциональных состояний, основа которых – в биологической чувствительности и активности. Тогда интенциональные состояния занимают свое онтологическое место рядом с qualia, становясь феноменами субъективной реальности, полностью не выразимыми в языке (как боль, цвет и т.п.). То есть я сначала на собственных состояниях учусь тому, что значит «знать», «полагать», «сомневаться», а потом уже по аналогии приписываю эти предикаты другим. В таком случае должно быть некоторое «состояние сомнения», в которое входит мозг и которое сопровождается «чувством сомнения», находящимся на стороне qualia. Причиной сомнения могут быть какие угодно обстоятельства внешнего мира, их следствием должно быть состояние сомнения, а его следствием – чувство сомнения. Полная аналогия с болью: я знаю, что такое боль, по своим ощущениям и догадываюсь, что другим бывает больно, по их поведению. Правда, Витгенштейн активно оспаривал такую концепцию боли[42], но дело даже не в этом. Боль – состояние не интенциональное: оно в себе содержит свой собственный предмет. А сомнение – всегда «сомнение в том, что...». Необходимо предполагается некоторый предмет сомнения и обстоятельство, которое делает его сомнительным. Возникает цепочка: (a) факт + фальсифицирующее его обстоятельство → (b) состояние мозга → (c) чувство сомнения.

Если следовать рассуждениям Сёрла, значением высказывания «я в сомнении» должны быть (b) или (c). Если мы говорим в первом лице, аналогия с болью может быть продолжена: я говорю «мое сомнение» и подразумеваю соответствующее состояние моего мозга или соответствующее «внутреннее» ощущение. Но что мы подразумеваем, когда говорим «его сомнение» или «он сомневается»? Мы делаем вероятностное умозаключение? Кто-то может сказать: но ведь когда мы говорим «у него бронхит», мы тоже подразумеваем воспаление бронхов, о котором догадываемся по внешним признакам. Значит, наши суждения об интенциональных состояниях других суть сплошь гипотезы?

Посмотрим, может быть сама идея ИИ поможет нам в концептуальном решении этой проблемы. Представьте себя на месте конструктора искусственного «мозга», перед которым стоит задача научить машину сомневаться. У вас есть комбинация (a), которую машине предстоит воспринять и оценить; вы не знаете, способны ли вы в принципе заставить ее испытывать (c) и насколько это вообще важно, а вот (b) – это как раз то, что и составляет вожделенное решение зада-

[42] *Wittgenstein L.* Op. cit. P. 257 and on.

чи. Предположим, что техническую задачу восприятия и идентификации машиной действительного положения дел вы решили. Вам необходимо получить от нее сигнал полной решительности (например, загорающуюся зеленую лампочку), если оцениваемые обстоятельства не оставляют места для сомнений, и сигнал сомнения (желтая лампочка), если имеющихся данных недостаточно для принятия определенного решения. Для большей реалистичности этого мысленного эксперимента предположим, что вы работаете в условиях ограниченного финансирования.

У вас может появиться искушение создать разветвленную нейронную сеть и моделировать в ней состояние человеческого мозга, соответствующее сомнению, всякий раз, когда к тому располагают оцениваемые обстоятельства, чтобы только через эту процедуру включалась желтая лампочка. Но в этом случае финансово ответственный участник вашей команды наверняка укажет вам на неоправданное расходование выделенных средств и будет прав. Да и зачем может быть нужно такое моделирование – чтобы воспроизвести (c)? Но вы никогда не узнаете, «почувствовала» ли машина свое чувство сомнения. Гораздо дешевле и разумнее, скажут вам, написать программу, основанную, например, на какой-либо «нечеткой» или «эпистемической» логике, которая позволит машине включать ту или иную лампочку в зависимости от вероятностной оценки предлагаемых обстоятельств.

Но тогда, возразите вы, точно так же можно имитировать и «болевое поведение», программно заставляя машину включать определенную лампочку при наличии определенного раздражителя. Более того, актер, изображающий испытывающего боль персонажа, сам ее скорее всего не испытывает... Хорошо, а если актер играет сомнение – в чем на самом деле состоит его актерская задача? Он должен правдоподобно имитировать «сомневающееся поведение». Значит ли это, что и сам он испытывает сомнение своего героя по ходу пьесы? Если нет, то бихевиористский принцип «сомнение есть сомневающееся поведение» оказывается под ударом.

Известно рассуждение Витгенштейна о «грамматике боли»: если бы не было «болевого поведения», то скорее всего было бы невозможно научить ребенка правильно использовать слово «боль»[43]. Можно ли научить кого-либо сомневаться, если не существует «сомневающегося поведения»?

Вернемся к «сомневающейся» машине. Предположим, что между ее рецепторами, воспринимающими сомнительную ситуацию (a), и «желтой лампочкой сомнения» мы поместили сложнейшую нейронную сеть, способную с максимальным приближением воспроизвести состояние человеческого мозга в момент сомнения. В каком случае

[43] *Wittgenshtein L.* Op. cit.

машина должна включить это состояние? Очевидно, когда ее рецепторы сообщат о положении дел, не позволяющем сделать желаемый вывод или принять желаемое решение с достаточно высокой степенью вероятности / обоснованности. И тогда на основе анализа входящих данных машина запустит программу, которая в свою очередь определит наличие оснований для сомнения и... переведет сложнейшую нейронную сеть в состояние, соответствующее «состоянию сомнения» человеческого головного мозга. А уже наличие этого состояния станет основанием для включения желтой лампочки. Не похоже ли это на странную конструкцию самолета, который после включения двигателей начинал бы хлопать крыльями, подражая взлету голубя?

Сторонник Сёрла может сказать, что включение нейронной сети в технологическую цепочку необходимо, если мы хотим моделировать именно сознательное сомнение, а не его автоматическую имитацию, поскольку именно ее состояние и есть необходимая онтологическая основа интенционального состояния, а следовательно, и сознания как такового. Но тогда резонен вопрос: если эта онтологическая основа сомнения на самом деле не необходима в случае с машиной (желтую лампочку – «сомневающееся поведение» – может включить сама управляющая программа), то почему мы решили, что она необходима сомневающемуся человеку?

Однако напрашивающееся в этом пункте возвращение к бихевиористскому тезису «сомнение есть сомневающееся поведение» уже заблокировано шахом, поставленным ему аргументом от актера, способного сыграть сомнение так же, как он играет боль, не испытывая ее на самом деле.

В действительности наша проблема в ее правильном концептуальном выражении состоит в другом: там, где один (человек) усомнится, другой будет слепо верить, а третий... тоже слепо верить, но в прямо противоположное и т.д. Если же мы построим два или более экземпляров сомневающейся машины, основываясь на простой технологии: рецептор → программа → лампочка, то они скорее всего будут завидным образом солидарны в выражении своих сомнений в сходных обстоятельствах. А ведь именно «свободу мышления», «творческий подход к оценке проблемного поля» и т.п. мы рассматриваем в наших онтологических презумпциях как сущностные характеристики собственно человеческого сознания в отличие от управляющих программ всевозможных автоматов. И вот в этом пункте сторонник «мышления мозгом» должен провозгласить: так вот за что ответственны сложнейшие комбинации нейронов головного мозга! У разных людей они хранят различные горизонты знаний и опыта, бессознательные комплексы, обусловливают их принадлежность к различным психоэмоциональным типам, и все это вместе предопределяет различные интенциональные состояния разных людей в одинаковых об-

стоятельствах. И поклонников этой в общем-то естественно-научной парадигмы объяснения не смущает такое концептуальное соображение: если нейронная модель помогает нам докопаться до причин индивидуальных различий в поведении и интенциональных состояниях, то... какая же это свобода? Свобода гражданина NN «сомневаться или не сомневаться» в моих глазах тогда состоит исключительно в моей досадной неосведомленности относительно сложного комплекса причин, которые в конечном счете предопределяют его выбор.

Но я не хотел бы здесь концентрировать внимание на защите свободы как фундаментальной человеческой ценности. В конце концов это не более чем еще одна онтологическая презумпция относительно человека. Гораздо интереснее следовать со всем возможным упорством за нашими основными путеводными ценностями: простотой и достаточностью концептуального воспроизводства возможных ситуаций.

Что сделает любой, даже не самый продвинутый программист, если его попросить построить два или более «сомневающихся» автомата, которые в сходных обстоятельствах выбирали бы различные интенциональные состояния? Правильно, в несложную программу, управляющую цветом лампочки, он добавит рандомизатор – генератор случайных значений. И тогда с высокой степенью вероятности там, где один автомат усомнится желтой лампочкой, другой зажжет зеленый свет. А теперь представим себе, что наличие другого с его (случайным?) интенциональным состоянием является важной составной частью проблемного поля первого автомата. То есть, его рецепторы воспринимают, а управляющая программа оценивает не только «верно, что *p*», но и «верно, что *p*, но *B* в этом сомневается». Предположим также, что между автоматами *A* и *B* существует некий языковой интерфейс и каждый из них предопределен к преследованию некоторой цели с использованием в том числе ресурсов другого и при этом способен к подбору средств ее достижения методом проб и самообучению на ошибках. Разве это не является достаточным описанием человеческой коммуникационной ситуации, в контексте которой только и имеют смысл такие интенциональные состояния, как «знание», «сомнение», «вера» и т.п.? Если это так и человеческий тип коммуникации в принципе воспроизводим машинами при указанных условиях, то интенциональные состояния – это не состояния мозга или осуществляющейся в нем программы. Более того, это и не модусы поведения, чем искушают возможные бихевиористские интерпретации. Это *модальности*, составляющие специфические *логические структуры* различных *коммуникационных ситуаций*.

Свободными в своей вере и своих сомнениях нас делают вовсе не нейронные сети, а коммуникационные ситуации.

Теперь самое время предложить ответы на вопросы, сформулированные вначале.

1. Может ли гипотеза ИИ внести свой вклад в традиционные философские дискуссии вокруг сознания и мышления? В чем, на мой взгляд, польза этой гипотезы?

1.1. Она избавляет формулировки проблемы от антропоморфности и психологизма и тем самым обнажает ее структурно-логический каркас.

1.2. Она подсказывает концептуально простое решение проблемы: слова «знание», «мышление», «сомнение» и т.п. обозначают не более и не менее чем взаимно определяемые позиции акторов в логической структуре той или иной коммуникационной ситуации.

1.3. Она ясно показывает, что мозг, компьютеры и другие естественные и искусственные устройства необходимы акторам для правилосообразного функционирования, но не достаточны для того, чтобы знать, мыслить или сомневаться.

Такую интерпретацию можно было бы обозначить как *коммуникативный функционализм*. Возможно, со временем обнаружится лучшее название, но пока остановимся на этом.

2. Может ли философия с помощью собственного неэмпирического инструментария предложить ответ на вопрос о возможности искусственного моделирования человеческого разума? – Если верен ответ на первый вопрос, то логично предположить, что наиболее перспективных исследований ИИ стоит ожидать в направлении создания мультиагентных систем.