КРИТИЧЕСКИЙ АНАЛИЗ ЛОГИКО-ЭПИСТЕМОЛОГИЧЕСКИХ ОСНОВАНИЙ ФИЛОСОФИИ ИСКУСТВЕННОГО ИНТЕЛЛЕКТА Х. ДРЕЙФУСА

В.А. Ладов

Проводится критический анализ философских воззрений Х. Дрейфуса на исследования в области искусственного интеллекта. Утверждается, что негативная оценка будущего технических интеллектуальных систем, высказанная Х. Дрейфусом, основывается на теоретической позиции, имеющей существенные затруднения логико-эпистемологического характера.

Ключевые слова: искусственный интеллект, теория, практика, формализация, логика, эпистемология, парадокс.

THE CRITICAL ANALYSIS OF LOGICAL AND EPISTEMOLOGICAL FOUNDATIONS IN ARTIFICIAL INTELLIGENT PHILOSOPHY OF H. DREYFUS

V.A. Ladov

The philosophical views of H. Dreyfus on researches in the area of artificial intelligent are analyzed in the article. It is asserted that theoretic position of H. Dreyfus has serious logical and epistemological problems. These problems are carefully considered in the article.

Keywords: artificial intelligent, theory, practice, formalization, logic, epistemology, paradox.

Хьюберт Дрейфус – современный американский философ, ставший известным, в первую очередь, за счет своих работ по феноменологии. Во многом благодаря его сочинениям американские философы, ориентированные, прежде всего, на аналитическую философскую традицию, стали знакомиться с гуссерлевской феноменологией. Однако не менее интересной и важной является и еще одна сторона творчества Х. Дрейфуса [1]. Он является одним из принципиальных критиков идеи искусственного интеллекта, опираясь в своих аргументах на феноменологическую традицию, а также на идеи позднего Л. Витгенштейна.

Х. Дрейфус полагает, что на основании основных результатов эпистемологических исследований в философии XX в., которые получены Э. Гуссерлем в поздней феноменологии «жизненного мира» [2], М. Хайдеггером в «Бытии и времени» [3], а также поздним Л. Витгенштейном в «Философских исследованиях» [4], можно утверждать приоритет практического, повседневного уровня бытия перед допускающими формализацию теоретическими конструкциями, и эта практическая деятельность человека в повседневном мире не просто сложна для формализации, но и вообще не позволяет провести подобные формально-логические процедуры. Основываясь на данном положении, Х. Дрейфус делает вывод, что исследования в области искусственного интеллекта, в рамках которых

осуществляется попытка логической формализации мышления человека с целью его имитации на искусственных технических носителях, обречена на провал, ибо наиболее фундаментальный практический слой бытия человека принципиально не формализуем.

Если двигаться в русле дрейфусовской мысли, то в ее подтверждение можно привести следующие пассажи из сочинений указанных выше философов. Так, Э. Гуссерль в работе «Кризис европейских наук и трансцендентальная феноменология» – сочинении, где представлена феноменология практического «жизненного мира» как поля наиболее фундаментальных очевидностей опыта сознания, пишет: «У человека в окружающем его мире есть много видов практики, и среди них этот своеобразный и исторически наиболее поздний: практика теоретическая» [2, с. 153]. Таким образом, во-первых, теория сама трактуется как особый вид практики жизненного мира, а во-вторых, этот вид практики признается наиболее внешним, производным по отношению к глубинным слоям практической жизни. В этом же произведении Э. Гуссерль дает нормативную установку будущим философским исследованиям: «Нужно полностью прояснить, т.е. привести к последней очевидности то, каким образом каждая очевидность объективно-логических свершений, дающая формальное и содержательное обоснование объективной теории (например, математической, естественнонаучной), имеет свои скрытые источники обоснования в той жизни, где осуществляются последние свершения, где очевидная данность жизненного мира всегда имеет свой донаучный бытийный смысл, где она приобрела его и вновь приобретает. От объективно-логической очевидности (от математического “усмотрения”, от естественнонаучного, позитивно-научного “усмотрения”, как оно выполняется ведущим свои исследования и дающим свои обоснования математиком и т.п.) путь здесь идет назад к праочевидности, в которой всегда заранее дан жизненный мир» [2, с. 175–176].

В работе М. Хайдеггера «Бытие и время» принципиальное значение приобретает дихотомия «наличное – подручное», которая оказывается аналогичной гуссерлевской дихотомии «теоретическое – практическое». Сфера подручного, с точки зрения М. Хайдеггера, имеет более глубокий онтологический смысл, нежели сфера наличного, представленная в качестве продукта научной теории: «Подручность есть онтологически-категориальное определение сущего как оно есть “по себе”» [3, с. 71]. Взгляд на сущее как на совокупность подручных вещей раскрывается в повседневной практической жизни: «Повседневное присутствие всегда уже есть этим способом, например, открывая дверь, я делаю употребление из дверной ручки» [3, с. 67].

Поздний Л. Витгенштейн выстраивает философско-лингвистическую концепцию, основным тезисом которой выступает тезис о принципиальной недоопределенности, незаконченности, нестабильности значения какого бы то ни было языкового выражения. Значение любого слова, в том числе и значение слова научного языка, определяется в контексте конкретной конечной лингвистической практики сообщества людей. Этот ход мысли Л. Витгенштейн противопоставляет традиционному представлению о научно-теоретическом мышлении и языке как об идеале объективности, строгости и однозначности. Данное представление признается философом не соответствующим реальному положению дел, поэтому он намеренно формулирует суждения, которые имеют явно провокационный характер по отношению к традиционным взглядам об идеале научности: «Ошибочно говорить, что есть что-то, что представляет собой значение…» [5, p. 3].

По отношению к указанному выше заявлению Х. Дрейфуса о том, что построение искусственного интеллекта невозможно ввиду принципиальной неформализуемости глубинного уровня человеческого мышления, можно выдвинуть критический аргумент непосредственно из лагеря сторонников искусственного интеллекта: о том, что между искусственным интеллектом и формально-теоретическим мышлением не следует ставить знак равенства. Есть попытки создания искусственного интеллекта на основе имитации неформального, практического уровня человеческого мышления. В частности, проект создания нейронных сетей.

Наша позиция по отношению к Дрейфусу тоже будет критической. Однако, со своей стороны, мы попытаемся представить иной путь рассуждений. Мы хотим подвергнуть критике позицию Дрейфуса, что называется, изнутри, используя так называемый аргумент от автореферентности. Мы хотим показать, что само рассуждение Дрейфуса содержит в себе определенный circulus vitiosus, если к нему самому применить его же собственные аргументы.

Не сложно заметить, что утверждение приоритета практического, повседневного бытия и мышления перед теоретическим в жизни человека производится Х. Дрейфусом на уровне теоретического мышления, с применением формально-логической аргументации. Возникает парадоксальная в эпистемологическом плане ситуация: более фундаментальный характер одной структуры обосновывается средствами другой структуры, которая признается второстепенной, производной по отношению к первой. Представляется, что данный эпистемологический парадокс может выступить основанием для сомнения в правомерности как дрейфусского рассуждения непосредственно, так и той позиции в рамках теоретической

философии XX в., которую Х. Дрейфус выбрал в качестве методологического инструментария для своей критики исследований в области искусственного интеллекта. И к Гуссерлю, и к Хайдеггеру, и к позднему Витгенштейну может быть применен аргумент от автореферентности, который покажет, что теоретическую позицию, в которой утверждается приоритет практического перед теоретическим, сложно назвать логически последовательной и эпистемологически реализуемой. Если философ релятивизирует научно-теоретический дискурс за счет его фундирования в локальных, нестабильных, разнообразных и несводимых друг к другу дискурсах человеческой практики, но при этом осуществляет сам акт данной релятивизации снова в теоретическом дискурсе, то он противоречит самому себе.

Но имеет ли теоретическую значимость сам аргумент от автореферентности, на котором мы строим свою критику? Можно ли его преодолеть, указав на несостоятельность подобного хода рассуждения?

В XX в., решая задачу преодоления теоретико-множественных и семантических парадоксов в философии математики и логике, были разработаны две хорошо известные концепции, критикующие идею автореферентности, поскольку именно она объявлялась основанием возникающих парадоксов. Это были семантическая теория А. Тарского [6] и теория типов Б. Рассела [7].

С точки зрения различения языка и метаязыка, проведенного А. Тарским в его семантической концепции, логические парадоксы, типа «Лжец» (а мы могли бы добавить, что и эпистемологический парадокс релятивиста, разновидностью которого является «парадокс Дрейфуса» в нашей интерпретации), возникают из-за неправомерного логико-лингвистического смешения в рассуждении, возникающего как раз на основании явления автореферентности. Неправомерной оказывается не позиция релятивиста, в которой утверждается, что истинность каких бы то ни было теоретических суждений релятивизируется относительно субъективных/интерсубъективных факторов познания (культурных, лингвистических, психических, биологических и т.д.), а как раз обвинение этой позиции в противоречивости. Считать высказывание «Все высказывания относительны» самопротиворечивым можно только исходя из ошибочного смешения различных уровней языка. На деле, само это высказывание относится уже не к языку, который в данном случае предстает объектом, о котором что-то говорится, а к метаязыку, и поэтому никакой противоречивости в утверждении релятивиста нет. Его высказывание «Все высказывания относительны» вполне может быть абсолютным, и это не приводит нас к некоему мыслительному коллапсу, если только мы

не забываем всякий раз проводить различия в уровнях языка, не допуская автореферентного дискурса.

С помощью теории типов Б. Рассела также может быть высказана критика в адрес тех, кто пытается уличить эпистемологические воззрения релятивизма в противоречивости. Подобно тем выводам, которые были сделаны из семантической концепции А. Тарского, можно сказать, что формулировка логического затруднения данных скептических воззрений основывается на смешении высказываний разных типов. Высказывание «Все высказывания относительны» попадает в логический тип более высокого порядка, нежели те высказывания, о которых в нем идет речь. Видимость противоречия возникает из-за неоправданного смешения данных логических типов, т.е. из-за формулировки автореферентных высказываний.

Однако указанные выше концепции, в которых установлен запрет на построение автореферентных высказываний, а значит, и аннулируется критический по отношению к Х. Дрейфусу аргумент от автореферентности, нам не представляются состоятельными. Они сами попадают в те же логические ловушки парадоксов, которые пытались преодолеть.

Так, теория типов Рассела, по сути, устанавливает запрет на универсалистский дискурс вообще. Нельзя говорить обо всем сразу, всегда следует помнить, что какое бы то ни было суждение может касаться только ограниченной предметной области. Следовательно, и истинностная оценка этого суждения также не может быть универсальной, она всегда должна релятивизироваться относительно того определенного круга предметов, который охватывается в суждении. Но как быть с самой формулировкой теории типов? Относится ли она сама только к определенному типу высказываний, охватывающих определенную, ограниченную предметную область, или все же представляет собой пример высказывания того самого универсального характера, запрет на которые она пытается установить? Формулируется ли сам принцип различения языка и метаязыка только еще в одном частном языке, по отношению к которому также возможна метапозиция, или же здесь используется некий универсальный язык, охватывающий собой все возможные лингвистические события? Когда Б. Рассел говорит о том, что общность классов в мире не может быть классом в том же самом смысле, в котором последние являются классами [8, с. 90], разве он не формулирует то свойство, посредством которого можно собрать в некую универсальную общность классов все возможные общности классов, а значит, и саму эту общность? Если это так, то сама формулировка теории типов представляет собой использование понятия класса всех классов, с которым она борется. Если это не так, то формулировка теории типов распространяется не на все

возможные общности классов, а только на некоторые, допуская возможность существования иных общностей, находящихся в метапозиции по отношению к ней и руководствующихся иным, отличным от теории типов, принципом отношения между классами. В итоге теория типов сама оказывается в логическом тупике. На подобные трудности в обосновании данной теории уже вскоре после ее появления указал П. Вайсс [9], представивший подробные критические аргументы по отношению к ней.

В отношении семантической концепции А. Тарского хорошо известна критическая аргументация Х. Патнема, в которой используется метафора так называемого языка красных чернил [10]. Если красными чернилами записываются правила для всех возможных языков, высказывания которых записаны чернилами всех иных известных цветов, то каким цветом будут записываться правила для языка красных чернил? Если красным, то сам этот язык оказывается замкнут на самом себе, т.е. автореферентным. Если же мы должны предположить существование чернил иного, неизвестного нам цвета, то правила языка красных чернил не будут распространяться на этот новый метаязык, и высказывания, записанные новым цветом, могут регулироваться иными правилами, отличными от разработанной семантической концепции.

То, что представленные в XX в. логико-семантические проекты не смогли аннулировать значимость аргументации, опирающейся на идею автореферентности, представляется нам крайне важным. На наш взгляд, идея автореферентности вообще является одной из определяющих для философии, фиксирующей сущность данного вида рациональной деятельности. В отличие от конкретных наук, которые ограничивают свои исследования определенным регионом сущего, философия всегда претендовала на то, чтобы быть универсальным знанием о сущем в целом. Собственно, в этом и состоит цель построения онтологической системы в философии – представить знание о сущем в целом на максимально высоком уровне общности. Выражение такого знания возможно именно в семантически замкнутом автореферентном языке, ибо только такой язык способен говорить обо всем, что есть, в том числе и о себе самом как определенном виде сущего. Эту специфику философского мышления подчеркивает Ф. Фитч: «Характерная черта философии состоит в том, чтобы дотянуться до этого максимального уровня и быть способной использовать автореферентные виды рассуждения, которые возможны на этом уровне» [11, p. 69].

Автореферентность имеет важнейшее значение для эпистемологии. Любая эпистемологическая концепция есть теоретическое построение о сущности, границах, нормах, идеалах и способах познания. При этом

само построение той или иной конкретной эпистемологической концепции есть проявление познавательных возможностей рационального субъекта. Для того чтобы предметная область исследования любой конкретной эпистемологической концепции была полной, она должна включать и построение самой этой концепции как один из вариантов проявления познавательных процессов. Только при исследовании полной предметной области можно говорить о допущении всеобщего и необходимого знания, знания как такового, отличного от мнения, которое всегда характеризуется ограниченностью. Установить полный запрет на автореферентность значит отказаться от концепта знания как такового в качестве регулятивной идеи познавательной деятельности.

Учитывая вышесказанное, представляется, что критические аргументы, высказанные в адрес философских воззрений Х. Дрейфуса, имеют важное значение, ибо показывают, что данная теоретическая разработка сначала сама должна разрешить собственные логико-эпистемологические затруднения, прежде чем настаивать на весомости своих критических положений по отношению к исследованиям в области искусственного интеллекта.

ЛИТЕРАТУРА

1. *Dreyfus H. L., Dreyfus S. E.* Making a Mind versus Modelling The Brain: Artificial Intelligence Back at a Branch-Point // The Philosophy of Artificial Intelligence / Boden M. (ed.) Oxford, 1990.

2. *Гуссерль Э.* Кризис европейских наук и трансцендентальная феноменология. СПб.: Владимир Даль, 2004.

3. *Хайдеггер М.* Бытие и время. М.: Ad Marginem, 1997.

4. *Витгенштейн Л.* Философские исследования // Витгенштейн Л. Философские работы. М.: Гнозис, 1994. Ч. 1. С. 76–319.

5. *Wittgenstein L.* Zettel. Oxford: Blackwell, 1967.

6. *Tarski A.* The Concept of Truth in Formalized Languages // Logic, Semantics, Metamathematics. Oxford: Oxford University Press, 1956. P. 152–278.

7. *Уайтхед А., Рассел Б.* Основания математики: в 3 т. Т. 1. Самара: Самар. ун-т, 2005.

8. *Рассел Б.* Философия логического атомизма. Томск: Водолей, 1999.

9. *Weiss P.* The Theory of Types // Mind. 1928. Vol. 37, № 147. P. 338–348.

10. *Патнем Х.* Реализм с человеческим лицом // Аналитическая философия: становление и развитие. М.: ДИК, 1998. С. 466–494.

11. *Fitch F.* Self-Reference in Philosophy // Mind. 1946. Vol. 55, № 217. P. 64–73.