А. С. Кляшторный
УрГУ им. А. М. Горького

КРИТИКА КОНЦЕПЦИИ «СИЛЬНОГО» ИИ В РАБОТЕ Х. ДРЕЙФУСА «ЧЕГО НЕ МОГУТ ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ?»

Первые основания идеи искусственного интеллекта (ИИ) обнаруживают себя еще в античных и средневековых источниках, в идее создания человекоподобных существ. Основаниями сущностного порядка становятся представления Аристотеля, Р. Луллия, Т. Гоббса, Г. В. Лейбница о возможности выделить правила мышления, о подобии процесса мышления вычислению. Фактическими основаниями ИИ как области научных исследований являются исследования К. Геделя, А. Тьюринга, Д. фон Неймана. Тест Тьюринга, гипотеза физической символической системы – одновременно главные ориентиры для исследований ИИ и основания концепции «сильного» ИИ начиная с 1950-х гг. Период 1950-70-х гг. в ИИ – период улучшения результатов с появлением более емких запоминающих устройств, более быстродействующих машин, с разработкой новых программ. В 1974 году, с открытием парадокса Моравека и эффекта «комбинаторного взрыва» наступает кризис – первая «зима» ИИ. Вышедшая в 1972 году книга Х. Дрейфуса «Чего не могут вычислительные машины?» стала не только предвестником грядущего кризиса, но и первым серьезным критическим замечанием в адрес исследований классического ИИ.

Дрейфус открывает свою работу кратким анализом истории развития философской мысли, предопределившей появление теории искусственного интеллекта. По мнению Дрейфуса, первым сторонником концепции синтаксического мышления, высказавшим мысль о том, что любое рассуждение может быть сведено к вычислению, был Сократ. Синтаксическая концепция мышления как процесса вычисления впервые была в явной форме сформулирована Т. Гоббсом. Задачу обнаружения первичных элементов – «словесных квантов» также безуспешно пытался решить Лейбниц. В XIX веке, подобно Гоббсу, считающий, что рассуждение есть вычисление, математик и логик Дж. Буль создает булеву алгебру – бинарную алгебру для представления элементарных логических функций. С появлением в 1835 году изобретений Ч. Бэббиджа теория начала

воплощаться в практику. Появление же в XX веке универсальной цифровой вычислительной машины, а позже и кибернетики, стали кульминацией философской традиции. Чуть позднее появился тест Тьюринга, ставший критерием проверки эффективности программы для вычислительной машины, проверкой ее интеллектуальности. Начался путь исследований в области искусственного интеллекта как исследований, стремящихся постичь самую суть рационального. Значительная часть исследований была направлена на формализацию поведения человека. К 1962-му году начала вырисовываться общая схема хода исследований: быстрый и эффектный успех, связанный с нетрудоемкими, простыми задачами, либо весьма неудовлетворительное решение сложных задач, затем снижение эффективности работы, разочарование, пессимизм.

Основным предположением, лежащим в основе исследований в области искусственного интеллекта, является предположение, согласно которому человек действует подобно устройству для символьной обработки информации. В свою очередь, данное предположение распадается на четыре допущения, которые Дрейфус последовательно опровергает:

1. Психологическое допущение. «Мышление можно рассматривать как переработку информации, заданной в бинарном коде, причем переработка происходит в соответствии с некоторыми формальными правилами» [1. С. 105].

2. Эпистемологическое допущение. «Все знания могут быть формализованы, то есть все, что может быть понято, может быть выражено в терминах логических отношений, точнее, в терминах булевых функций – логического исчисления, задающего правила обращения с информацией, заданной в двоичном коде» [1. С. 105].

3. Онтологическое допущение. «Машинная модель мышления предполагает, что все сведения о мире, все, что составляет основу разумного поведения, должно в принципе допускать анализ в терминах множества элементов, безразличных к ситуациям» [1. С. 105]. Таким образом, все происходящее в мире можно представить в виде множества фактов, каждый из которых логически не зависит от остальных.

4. Биологическое допущение. «На некотором уровне – обычно полагают, что на уровне нейронов – операции по переработке информации носят дискретный характер и происходят на основе

некоторого биологического эквивалента переключательных схем» [1. С. 105].

Психологическое допущение оказывается возможным лишь на основании смешения обычного смысла слова «информация со специальным значением, имеющим место в кибернетической теории информации.

Опровержение психологического допущения, однако, нисколько не мешает продолжать существовать допущению эпистемологическому. В данном случае имеет место неоправданное распространение методологии естественных наук на область психических явлений. Однако, пишет Дрейфус, «полное опровержение эпистемологического допущения потребовало бы доказательства того, что мир принципиально не может быть проанализирован в терминах четко определенных данных» [1. С. 164], а подобное доказательство одновременно является опровержением онтологического допущения. По мнению Дрейфуса, онтологическое допущение в большей степени не согласуется с нашим опытом, Оно явилось следствием необходимости понимания мира и умения управлять им, принуждавшим западную традицию к упрощению действительности, в то время как она гораздо сложнее. Распространению этого стремления к упрощению способствовали успехи в физике. Однако, хотя в рамках физической теории мир и может быть представлен в виде совокупности множества атомарных фактов, однако, данное представление, выходя за пределы данной теории, плохо согласуется с нашим опытом. То, что мир может быть «разделен» на атомарные факты, еще не означает, что, попытавшись «собрать» из этих элементов целостную картину мира, мы действительно ее получим.

В рамках биологического же допущения лежит интерпретация нейронного импульса в качестве единицы информации, циркулирующей в мозгу, подобно машинному биту. Возможность подобной интерпретации, однако, опровергается Дрейфусом, что отменяет и биологическое допущение.

В силу того, что опровергнутые Дрейфусом допущения лежали в основании предположения о том, что человек действует подобно устройству для символьной обработки информации, то опровергнутым оказывается как данное предположение, так и вся концепция классического ИИ, для которого оно является фундаментальным. Все те серьезные трудности, которые встали

перед разработками в области ИИ, уже не могут быть решены путем увеличения быстродействия и объемов памяти. Опровержение же вышеуказанного фундаментального предположения ИИ, не оставляет никаких оснований для уверенности в возможности моделирования человеческого поведения. А в силу этого излишним будет даже само появление вопроса о возможности прохождения машиной теста Тьюринга.

Завершает свою работу Дрейфус предположением о том, какими последствиями чревато подобное плачевное положение исследований в области ИИ, теоретиков которого сам он именует не иначе как «последними метафизиками»: «Говоря словами М. Хайдеггера, если западная метафизика достигла в кибернетике своей высшей точки, то возникшие за последнее время трудности в разработке проблем «искусственного интеллекта» отражают не столько недостаточное развитие нашей технологии, сколько, пожалуй, указывают на принципиальные границы ее возможностей» [1. С. 191].

Работа Х. Дрейфуса «Чего не могут машины» стала первой философской работой, в которой была подвергнута критике идея о том, что человек подобен устройству символьной обработки информации, идея, являющаяся фундаментальным основанием концепции сильного ИИ в целом. Написание именно этой работы сделало возможным появление в будущем знаменитой статьи Д. Серля «Разум, мозг и программы» и его примера «Китайская комната», ставшего, возможно, «надгробным камнем» на «могиле» концепции «сильного» ИИ.

Примечания

1. *Дрейфус Х.* Чего не могут вычислительные машины. Критика искусственного разума. / Пер. с англ. Н. Родмана. Под ред. Б. В. Бирюкова. М.: Прогресс, 1978.