Секция «История зарубежной философии»

Критика искусственного интеллекта в философии Хьюберта Дрейфуса

Научный руководитель – Кузнецов Антон Викторович

Власова Дарья Александровна
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Философский факультет, Кафедра истории зарубежной философии, Москва, Россия
E-mail: daalz@mail.ru

Хьюберт Дрейфус (1929 - 2017 гг.) - философ, который в своих идеях своеобразно объединил аналитическую и континентальную линии философии. В ключевых работах на тему критики ИИ «Alchemy and Artificial Intelligence» (1965 г.) [4], «Чего не могут вычислительные машины: Критика искусственного разума» (1972 г.) [1], «Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer» (1986 г.) [2], «What Computers Still Can't Do: A Critique of Artificial Reason» (1992 г.) [3] Х. Дрейфус представил пессимистическую оценку развития искусственного интеллекта и ряд ключевых аргументов против возможности создания ИИ на базе существующих компьютерных технологий.

Его первая серьезная работа, посвященная критике искусственного интеллекта - это отчет для компании RAND, куда Хьюберта наняли в качестве консультанта для независимой оценки (в виде философского анализа) проекта Саймона, Ньюэлла и Шоу Cognitive Simulation, который был назван «Алхимия и искусственный интеллект» [4]. В отчете Дрейфус жестко расритиковал проект, сравнив новую область исследования с алхимией. Х. Дрейфуса в большей степени «зацепила» уверенность разработчиков в том, что они стоят на пороге создания аналога человеческого разума и что, достигнув этого, они смогут разгадать все тайны мышления человека. В этом отчете появляется центральный аргумент Х. Дрейфуса против разработчиков программ ИИ. Философ отстаивает позицию, что программисты и разработчики принимают необоснованное допущение: человеческое сознание можно свести к формальному оперированию символами, так возникает уверенность, что в программе можно воплотить все те операции, которые осуществляет человек в процессе мышления (ассоцианистское допущение). Х. Дрейфус выступает резко против тенденции к редуктивному пониманию человеческого способа мышления.

В дальнейшем он расширит и углубит свои идеи, в работе «Чего не могут вычислительные машины: Критика искусственного разума» (1972 г.) он отразил все свои основные замечания и претензии к программистам [1]. Он сформулировал четыре основных допущения, которые, по мнению философа, совершенно необоснованно принимают разработчики программ ИИ:

1. Биологическое допущение - упрощение работы сознания. Представление о том, что на уровне нейронов процесс передачи информации носит дискретный характер и происходит по довольно определенной и простой схеме, следовательно, этот процесс можно воплотить на не биологическом носителе.

2. Психологическое допущение - упрощение тех функций, которые осуществляет наше сознание. Представление о том, что наше мышление - это процесс обработки информации согласно определенным формальным правилам.

3. Эпистемологическое допущение - вся информация, которую «обрабатывает» человеческое сознание, может быть формализована и выражена в терминах логических отношений, точнее, в терминах булевых функций - логического исчисления, задающего правила обращения с информацией, заданной в двоичном коде.

4. Онтологическое допущение - вообще все, что может быть существенно для разумного поведения, может быть представлено в дискретных, четко определенных терминах.

В книге «Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer» (1986 г.) Хьюберт Дрейфус вместе со своим братом Стюартом исследуют вопросы соотношения вычислительного и интуитивного типов рациональности [2]. Изучая разные случаи применения логики и интуиции на практике, они приходят к выводу о том, что интуитивный способ взаимодействия с миром, который основан на жизненном опыте человека, чаще становится более эффективным, нежели вычисляющая рациональность, основанная на ограниченных фактах, теоретических моделях и логике, которые используют в разработке ИИ. Философ делает акцент на том, что практическое действие и опыт не поддаются формализации, нельзя объяснить, как ездить на велосипеде, этому можно только научить с помощью практических занятий.

Однако, разработки в области создания программ искусственного интеллекта не стояли на месте, и те «действия», которые раньше казались невозможным для ИИ, с точки зрения Х. Дрейфуса, стали обыденной реальность в среде программистов. Многие из критических замечаний и ограничений, которые выдвинул Х. Дрейфус в работе 1972 года были преодолены. Это требовало от философа некоторого пересмотра своих наиболее резких формулировок и замечаний. Тем не менее в своей более поздней работе, посвященной критике ИИ «What Computers Still Can't Do: A Critique of Artificial Reason» (1992 г.), Х. Дрейфус мало что изменил [3]. Он написал введение к переизданию книги, где обозначил свою позицию, что в целом его предположение о том, что разработчики принимают необоснованные допущения, оказалось верным. ИИ не вышел за рамки его критики в своей основе, тем не менее, учась на своих ошибках, разработчики программ ИИ смогли совершить достаточно серьезный шаг вперед и решить ряд сложных задач, что все же не приблизило их к созданию аналога человеческого разума потому, что эти вещи совершенно отличны по своей природе и представляют два разных типа рациональности.

В ходе исследования мы увидели, что общефилософский интерес и жизненный путь приводят Х. Дрейфуса к критике искусственного интеллекта и высоких технологий сквозь призму континентальной философии, что совсем не свойственно американским философам. Его позиция крайне интересна и актуальна. На сегодняшний день мы можем проследить те моменты, где Х. Дрейфус был слишком строг к ИИ, а где его замечания оказались адекватными современной ситуации. Критика ИИ со стороны Х. Дрейфуса во многом философская, но при этом становится понятно, что философ был достаточно глубоко погружен в технические вопросы создания программ ИИ, что и позволяло ему говорить, что ИИ не равен и не будет равен естественному интеллекту. Однако в рамках собственной оценки идей философа был сделан вывод, что Х. Дрейфус считает невозможным создание ИИ, который воплотит в себе разум человека, но на самом деле все его аргументы могут сказать нам только о том, что ИИ действует не так как человек, что в принципе не отрицают программисты, хотя и пытаются сравнивать ИИ и разум человека.

Источники и литература

1) Дрейфус Х. Чего не могут вычислительные машины. М.: Прогресс, 1978.
2) Dreyfus H. L., Dreyfus S. E. Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer. N.Y.: The Free Press, 1986.
3) Dreyfus H. L. What Computers Still Can't Do: A Critique of Artificial Reason. Cambridge, MA: MIT Press, 1992.
4) 4. Dreyfus H. L. Alchemy and Artificial Intelligence. RAND papers, 1965. [Электронный ресурс] // Режим доступа: http://rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf (дата обращения 08.02.21).