УДК 11+004.8+17+51-7

Робофилософия: философствование людей о роботах, или философствование роботов о мире, в котором они существуют?

© | Лобовиков Владимир Олегович vlobovikov@mail.ru

ИФиП УрО РАН, Екатеринбург, Россия

Аннотация. *Рассмотрено множество различных значений слова «робофилософия». Приведено множество определений понятий, составляющих основание для конструирования математической модели философии, могущей быть использованной для конструирования собственно философской подсистемы «сильного» ИИ автономного универсального робота. Сделаны выводы, что: для «слабого» искусственного интеллекта, морально-правовая и метафизическая подсистемы излишни (не нужны); для гипотетического «сильного» искусственного интеллекта автономного универсального робота, ценностные подсистемы (в частности, морально-правовая и метафизическая) необходимы.*

Ключевые слова: *гипотетический сильный искусственный интеллект, автономный робот, эмоциональный робот, ценности, метафизическая подсистема искусственного интеллекта робота, дискретная математическая модель философии, алгебра метафизики как формальной аксиологии*

Robot - Philosophy: Human's Philosophizing about Robots, or Robot's Philosophizing of the World they Exist in?

V.O. Lobovikov

IPAL UB RAS, Yekaterinburg, Russia

Abstract. A set of different meanings of the word "robot-philosophy" is considered. A set of definitions is given for the notions, making up grounds for constructing a mathematical model of philosophy which could be utilized for constructing a proper philosophical subsystem of "proper" AI of an autonomous universal robot. The following conclusions are made: for the "weak" AI, moral-legal and metaphysical subsystems are redundant (unnecessary); for the hypothetical "strong (proper)" AI of an autonomous universal robot, axiological (evaluative) subsystems (in particular, moral-legal and metaphysical ones) are necessary.

Keywords: hypothetical-artificial-intelligence-proper, autonomous-robot, emotional-robot, values, metaphysical-subsystem-of-artificial-intelligence-of-robot, discrete-mathematical-representation-of-philosophy, algebra-of-metaphysics-as-formal-axiology

В течение многих веков обычные, т. е. *статистически* нормальные, гуманитарии свято верили (и до сих пор продолжают верить) в то, что «алгеброй гармонию поверить» невозможно в принципе. Это их кредо (символ веры)

допускает использование *метафоры* «алгебра гармонии» (как именно метафоры) в *художественных* произведениях, но исключает возможность использования этого словосочетания в буквальном смысле в серьезных научных работах. Согласно позитивистам, генерируемые гуманитариями тексты по метафизике, аксиологии, и теологии семантически бессмысленны и, поэтому, подлежат устранению путем логического анализа естественного языка [1–5]. Следовательно, философия и аксиология совершенно неуместны также и в искусственном интеллекте роботов, так как, согласно классическому позитивизму, очевидно, что «робофилософия» есть нонсенс, ибо любая философия — нонсенс. С некоторой точки зрения, в идеале, ИИ-робот есть адекватная *модель* человека. По мнению позитивистов, для человека (выступающего в качестве *оригинала* для этой модели), метафизика не нужна и даже вредна. Поэтому, с точки зрения позитивистов, она не нужна и для ИИ-роботов. В отношении роботов с так называемым «слабым» интеллектом позитивисты, на мой взгляд, правы: для не имеющего свободы (привинченного к конвейеру) жестко запрограммированного узко специализированного производственного ИИ-робота философия не нужна и даже вредна (мешает эффективной работе).

Однако, вообще говоря, позитивистское философствование об ИИ-роботах, на мой взгляд, неадекватно и в теоретическом отношении, и в отношении прогрессивного развития человеческой практики, особенно, в отношении универсальных автономных роботов, обладающих так называемым «сильным» ИИ. В настоящее время статус таких роботов является гипотетическим. Но на уровне фундаментальных научных исследований эта нетривиальная гипотеза продолжает разрабатываться; она отнюдь не исключается собственно научным подходом к стратегически важной для перспектив человечества комплексной междисциплинарной научно-технической программе, именуемой «искусственный интеллект».

Но что такое искусственный интеллект и интеллект вообще? Удовлетворительного общепринятого определения значения слова «интеллект» в научной литературе пока нет. Но в таком случае научное обсуждение искусственного интеллекта оказывается чрезмерно неопределенным. Поэтому, чтобы снизить неопределенность предмета дискуссии до разумных пределов и обеспечить возможность точного понимания читателем данной работы, примем в ней в качестве «рабочей гипотезы», фиксирующей *необходимый* аспект пока еще отсутствующего безупречного определения интеллекта (вообще), следующее суждение.

Интеллект вообще, т. е. как естественный, так и искусственный, согласно принимаемой в настоящей работе гипотетической дефиниции, есть сложная система, *необходимо* состоящая из следующих подсистем: логико-научная; морально-правовая; художественно-эстетическая; религиозно-метафизическая. Конечно же, такое «определение» нельзя считать удовлетворительной общепринятой дефиницией, так как, например, позитивисты ее не примут из принципа. Тем не менее, в настоящей работе эта спорная «дефиниция» принимается

(в качестве «рабочей гипотезы») и тем самым неопределенность предмета дискуссии об ИИ-роботах существенно снижается. Следствием принятия такой спорной дефиниции оказывается осознание того факта, что в подавляющем большинстве случаев современные «роботы с искусственным интеллектом» суть адекватные модели людей-инвалидов особого рода, а именно, людей-олигофренов, т. е. людей с тяжелыми повреждениями интеллекта.

То, что некоторые интеллектуальные результаты и возможности некоторых современных «роботов с искусственным интеллектом» существенно превосходят соответствующие результаты и возможности нормальных людей, объясняется тем, что такие «интеллектуально превосходящие людей» роботы суть адекватные модели таких людей-олигофренов, для которых существует специальное название «олигофрен-савант». Такие люди-олигофрены значительно превосходят нормальных людей по своим возможностям и результатам решения некоторого типа специальных задач, но в целом, являются инвалидами с тяжелыми повреждениями интеллекта. У одних олигофренов-савантов может вообще отсутствовать (или быть в зачаточном состоянии) художественно-эстетическая подсистема интеллекта, у других — морально-правовая (или и то и другое сразу). В случае отсутствия или тяжелого повреждения *ценностных* подсистем интеллекта, он полностью (или почти полностью) сводится к его логико-научной подсистеме, а это может создавать или некий психологический дискомфорт, или даже большую опасность для общества.

Это тема очень важная и очень сложная. Из-за ее большой сложности, в настоящей статье целесообразно сосредоточить внимание лишь на некоем вполне определенном небольшом фрагменте указанной темы, а именно, лишь на собственно *метафизической* подсистеме искусственного интеллекта. И. Кант был крайне удивлен и искренне возмущен тем явно неудовлетворительным состоянием, в котором находилась в его время метафизика, и систематически критиковал ее, но считал *необходимой* для человеческого интеллекта [6, 7]. Если метафизика — *необходимая* подсистема любого интеллекта вообще, то, изобретая ИИ-роботов будущего, как представить метафизические знания людей в «сильном» искусственном интеллекте будущих *универсальных* автономных роботов? Естественно предположить, что для адекватного представления по крайней мере некоторой важной части таких знаний в ИИ потребуется некий современный вариант того, что Г.В. Лейбниц называл "characteristica universalis" [8–10]. Иначе говоря, потребуется создание и развитие такого нового направления научных исследований, которое может быть названо «вычислительной философией».

В отношении словосочетания «вычислительная философия», явно бросающего дерзкий вызов парадигме, традиционно господствовавшей в истории философии (Р. Луллий и Г.В. Лейбниц были очень странными и очень редкими «исключениями из правила»), здесь уместно заметить, что, согласно фундаментальным результатам А. Тьюринга, А. Черча, А.А. Маркова (младшего), Э.Л. Поста, С.К. Клини, А.Н. Колмогорова и других выдающихся математиков XX века, вообще говоря, в некотором смысле, *не все эффективно вычислимо*.

Следовательно, область адекватности вычислительного аспекта рационалистического оптимизма Лейбница должна быть сознательно ограничена.

Глубокий философско-математический анализ проблем «вычислительной философии» всесторонне представлен, например, в монографии Ю.А. Ершова и В.В. Целищева [11]. В настоящей статье их «отрезвляющая» (сдерживающая чрезмерный «вычислительный оптимизм» в философии) позиция полностью принимается как безупречно обоснованный научный взгляд на современную ситуацию в философии математики (и на попытки плодотворного приложения математики к философии). Однако, осознание ограниченности возможностей реализации сформулированной Лейбницем идеи «вычислительной философии» не означает отказ от систематического изучения и использования таких возможностей в пределах той *некустой* сферы, в которой «вычислительная философия» вполне адекватна и плодотворна.

Для исключения возможных концептуальных недоразумений, целесообразно прояснить значение слова «метафизика», фактически имеющего в естественном языке философии не одно единственное, а несколько различных значений. Так, например, в диалектико-материалистической философии (популярное сокращенное название — «диамат»), господствовавшей в СССР в течение многих десятилетий, слово «метафизика» обозначало некое ложное и, поэтому, вредное мировоззрение, отрицающее существование всеобщей взаимосвязи в материальном мире, а также отвергающее его внутреннюю противоречивость. Согласно диалектическому материализму, в настоящее время потерявшему доминирующее положение в отечественной философии, «метафизика» есть противоположность (отрицание) «диалектики», утверждающей внутреннюю противоречивость материального мира и существование в нем всеобщей взаимосвязи.

Аналогичную (сходную) отрицательную позицию по отношению к метафизике в течение столетий занимали и продолжают занимать позитивисты. Они считали и продолжают считают метафизику бессмысленной и, поэтому, вредной. Таким образом, на первый взгляд, позитивисты и марксисты-ленинцы едины в своем негативном отношении к метафизике, однако, внимательно читая и анализируя соответствующие тексты, можно заметить, что представители указанных двух философских течений используют слово «метафизика» отнюдь не в одном и том же значении. В отличие от марксистов-ленинцев, используя слова «философия» и «метафизика» как синонимы, позитивисты отвергают любую философию, в том числе и диалектико-материалистическую (философию марксизма-ленинизма). В данной статье, в отличие от марксистов-ленинцев, слова «философия» и «метафизика» используются как синонимы. Но, в отличие от позитивистов, метафизика (=философия) рассматривается в данной работе как вполне осмысленная интеллектуальная дисциплина, отказ от которой чреват инвалидностью (серьезным поражением) интеллекта. Механической работе на конвейере такая инвалидность не мешает: такой человек-инвалид вполне может выполнять полезную для общества работу; он в каких-то конкретных отношениях доста-

точно адаптивен и социализируем, но такое серьезное поражение интеллекта несовместимо с предложенной выше дефиницией нормального интеллекта вообще (как естественного, так и искусственного), а, следовательно, несовместимо и с гипотетической идеей «сильного» искусственного интеллекта автономного универсального робота.

Диссидентская (как с точки зрения марксизма-ленинизма, так и с точки зрения позитивизма) идея истолкования *общей* философии как некой *абстрактной формальной аксиологии*, а таких *частных* философских дисциплин как этика, эстетика, философия права, в качестве конкретных примеров или проявлений (разделов) этой гипотетической абстрактной формальной аксиологии, возникла у автора настоящей статьи в самом начале семидесятых годов прошлого века, но в публикации рукописей статей по алгебре формальной этики и алгебре формальной эстетики автору тогда было отказано из принципа, так как *марксизм-ленинизм боролся с формализмом* в искусстве, науке, философии, т. п. Первые тщательно замаскированные автором публикации такого рода появились лишь в самом конце семидесятых и начале восьмидесятых годов прошлого века. Вначале, вместо «политически некорректного» термина «(не классовая, а общечеловеческая) алгебра *формальной* этики» использовался более «нейтральный» термин «алгебра поступков». «Идеологически неприемлемые» термины «алгебра *формальной* этики» и «алгебра *естественного* права» стали использоваться позднее. Систематическое обсуждение автором морально-правового аспекта комплексной междисциплинарной научно-технической программы «Искусственный интеллект» вообще, и проблемы морально-правового программирования деятельности автономных ИИ-роботов в особенности, началось с середины восьмидесятых годов прошлого века [12].

Собственно метафизическая (философско-онтологическая и эпистемологическая) подсистема искусственного интеллекта автономного универсального ИИ-робота стала обсуждаться позднее. Термин «алгебра метафизики (как формальной аксиологии)» начал появляться в работах автора с начала XXI века [13–16]. Часто (иногда с дипломатической целью, а именно, для частичной маскировки концепции, бросающей вызов не только марксизму-ленинизму, но и позитивизму, который, по-прежнему, весьма влиятелен и многим кажется интеллектуально респектабельной доктриной), вместо термина «алгебра метафизики (как формальной аксиологии)» используется термин «алгебра философии (как формальной аксиологии)» или «алгебра формальной аксиологии)» [13–16]. Примерно с 2015 года по настоящее время автором конструируются *логически формализованные* мультимодальные *аксиоматические системы* «вычислительной философии» (для ИИ-роботов), синтезирующие собственно-философскую онтологию, универсальную эпистемологию и формальную аксиологию). Для точного определения семантики искусственных языков этих формальных аксиоматических теорий используется алгебра метафизики как формальной аксиологии.

Тот очевидный факт, что на очень многих простых людей (и особенно на позитивистов) тексты и речь философов (=метафизиков) часто производят

впечатление бессмысленного потока слов сумасшедшего, можно объяснить тем, что очень часто имеет место логико-лингвистическое недоразумение. Дело в том, что, в сущности, *тексты метафизиков имеют формально-аксиологическую семантику*, а люди, безуспешно пытающиеся понять эти тексты, безуспешно пытаются найти их дескриптивно-индикативную семантику. Однако, отсутствие дескриптивно-индикативных значений у слов и словосочетаний не означает отсутствие у них семантических значений вообще. Слова и словосочетания естественного языка философии могут иметь и, как правило, имеют абстрактные *ценностные* (формально-*аксиологические*) значения [16]. Вполне естественно, поэтому, что, отвергая аксиологию, позитивисты *неизбежно* отвергают тем самым и метафизику. Итак, в основании настоящей статьи лежит весьма нетривиальная презумпция, согласно которой, в сущности, *метафизика* (=*философия*) *есть формальная аксиология*. Если к понимаемой таким образом метафизике приложить некую подходящую для этого алгебру, то в результате такого приложения математики к философии можно получить некую алгебру метафизики. Слово «алгебра» используется здесь в собственно математическом значении. Но более точно и строго говоря, речь в данной статье идет не просто об алгебре, а о двузначной *алгебраической системе* метафизики как формальной аксиологии (алгебра — «вырожденный» случай алгебраической системы [17]).

Двузначная алгебраическая система метафизики как формальной аксиологии есть тройка множеств $< \Phi, O, R >$, в которой символ Φ обозначает непустое (даже бесконечное) множество всех таких и только таких объектов (существующих или несуществующих, — неважно), которые являются либо хорошими либо плохими, с точки зрения некого *оценщика* (субъекта оценивания) Σ (индивидуального или коллективного, естественного или искусственного, — неважно). Очевидно, что Σ — *переменная*: изменение ее значений может приводить к изменению оценок конкретных элементов множества Φ. Однако, если значение переменной Σ определено (зафиксировано), то оценки конкретных элементов множества Φ оказываются вполне определенными. Элементы множества Φ будем называть *формально-аксиологическими объектами* метафизики независимо от их бытия или небытия. Символы «+ (хорошо)» и «– (плохо)» обозначают *абстрактные аксиологические (ценностные) значения* элементов множества Φ. Моральные поступки, или юридические акты, или лица (индивидуальные или коллективные, естественные или искусственные, — неважно) суть *конкретные примеры* (*частные случаи*) элементов множества Φ.

В тройке множеств $< \Phi, O, R >$, символ O обозначает множество всех n-*арных алгебраических операций* (или просто *операций*), определенных на множестве Φ. Элементы множества O называются *формально-аксиологическими* операциями двузначной *алгебры метафизики* (как формальной аксиологии). В двузначной алгебраической системе формальной аксиологии $< \Phi, O, R >$, определенные на множестве Φ n-*арные алгебраические операции* являются такими и только такими *функциями*, которые ставят в однозначное

соответствие каждой упорядоченной n-ке элементов множества Φ некоторый элемент множества Φ, называемый результатом применения упомянутой n-арной алгебраической операции к упомянутой упорядоченной n-ке элементов множества Φ. Иначе говоря, определенная на множестве Φ n-арная (*алгебраическая*) *операция* есть n-местная функция $f: \Phi^n \to \Phi$. Итак, областью допустимых значений тех (именуемых алгебраическими операциями) функций, которые суть элементы множества O, является множество Φ; и областью изменения значений этих функций (алгебраических операций, определенных на Φ) служит то же самое множество Φ. Конкретные примеры унарных и бинарных алгебраических операций, принадлежащих множеству O, читатель может найти в [12–16].

Определим теперь значение символа R в упомянутой выше тройке < Φ, O, R >. В этой тройке, символ R обозначает множество всех n-*местных формально-аксиологических отношений*, определенных на множестве Φ. Таковым является, например, определенное в [12–16] бинарное отношение «*формально-аксиологическая эквивалентность* (элементов множества Φ)».

Результаты алгебраических операций, определенных на множестве Φ, суть элементы множества Φ, следовательно, они или хороши или плохи (с точки зрения Σ). Между аксиологическими (ценностными) значениями (х или п) тех элементов множества Φ, к которым применена формально-аксиологическая алгебраическая операция, определенная на множестве Φ, и аксиологическим значением (х или п) результата этой операции существует ценностно-функциональная связь. Ценностное значение (х или п) результата определенной на Φ алгебраической операции есть значение некой *ценностной функции*, допустимыми значениями переменных которой являются аксиологические значения (х или п) тех элементов множества Φ, к которым применена упомянутая алгебраическая операция.

По определению, *ценностной функцией* называется такая и только такая функция, у которой областью допустимых значений ее переменных является двухэлементное множество {х (хорошо), п (плохо)}, и областью (изменения) значений этой функции является то же самое двухэлементного множество. Иначе говоря, при рассуждениях о ценностных функциях, имеются в виду следующие отображения: $\{х, п\} \to \{х, п\}$, если разговор идет о функциях, детерминируемых *одним* ценностным аргументом; $\{х, п\} \times \{х, п\} \to \{х, п\}$, если разговор идет о функциях, детерминируемых *двумя* ценностными аргументами (здесь «×« обозначает Декартово произведение множеств); $\{х, п\}^N \to \{х, п\}$, если разговор идет о функциях, детерминируемых N ценностными аргументами, (здесь N обозначает некое конечное положительное целое число).

С помощью аккуратного «вычисления» соответствующих композиций ценностных функций в двузначной алгебраической системе метафизики, не только естественный интеллект человека, но также и искусственный интеллект автономного универсального ИИ-робота может генерировать конечные

(но потенциально бесконечные) системы формально-аксиологических «уравнений», адекватные переводы которых на естественный язык людей оказываются нетривиальными метафизическими принципами, открытыми, уточненными и обоснованными в процессе многовекового развития содержательной человеческой философии на уровне естественного языка существ вида homo sapiens. Двузначная алгебраическая система метафизики как формальной аксиологии логически непротиворечиво синтезирует в себе многие важные мировоззренческие открытия и изобретения человечества, например, основные положения философских доктрин Парменида и Гераклита, которые, на первый взгляд, кажутся логически несовместимыми, но, используя упомянутую алгебраическую систему метафизики, ИИ может самостоятельно генерировать все основные парадоксальные афоризмы элеатов и, как это ни странно, даже основные «темные» сентенции Гераклита Темного. О таком удивительном синтезе противоположностей см. [13, 15].

Читатели, изучавшие математическую логику, могут заметить удивительное *сходство* (подобие) между определенной выше двузначной алгебраической системой формальной аксиологии и двузначной алгебраической системой классической формальной логики. Это их *сходство* (подобие) поразительно и очень важно: оно делает эти две *качественно различные* системы *аналогами* (*моделями*) друг друга. Однако, *отношение сходства* (*подобия*) *не является отношением тождества*. Последнее (отношение) есть отношение эквивалентности, а первое им не является, так как, вообще говоря, сходство (подобие) не является транзитивным (отношением).

На основании вышесказанного естественно предположить, что системообразующим фактором для любого интеллекта как сложной системы, необходимо включающей в себя все вышеупомянутые качественно различные подсистемы, является отношение существенной аналогии между ними (т. е. их существенного подобия и качественного различия), — предпосылка для взаимного моделирования этими подсистемами друг друга. Подразумевается, что такое отношение эвристически плодотворного взаимного моделирования имеет место в любой нормальной, (т. е. полноценной, неповрежденной «Бритвой Оккама») интеллектуальной системе. Приводящее к инвалидности «обрезание» кажущегося «лишним» в интеллекте (например, его метафизической или художественно-эстетической подсистем) ведет к существенному снижению эвристических потенций интеллекта, т. е. к его развитию путем регресса.

Литература

[1] Айер А.Дж. *Язык, истина и логика*. Москва, Канон+, Реабилитация, 2010, 240 с.
[2] Карнап Р., Гэн Г., Нейрат О. *Научное миропонимание – Венский кружок*. Erlenmais («*Познание*»), *Избранное*. Москва, Территория будущего, Идея-Пресс, 2006, с. 57–74.
[3] Поппер К. *Логика и рост научного знания*. Москва, Прогресс, 1983, 605 с.
[4] Шлик М.О *Фундамент познания. Аналитическая философия. Избранные тексты*. Москва, МГУ, 1993, с. 33–50.

[5] Carnap R. The Rejection of Metaphysics. 20th-Century Philosophy: The Analytic Tradition. New York: Free Press, 1935, pp. 206–220.

[6] Кант И. *Пролегомены ко всякой будущей метафизике, могущей возникнуть в смысле науки*. Москва, Академический проект, 2008, 174 с.

[7] Кант И. *Критика чистого разума*. Москва, Эксмо, 2012, 736 с.

[8] Лейбниц Г.В. *Об универсальной науке, или философском исчислении. Сочинения в 4 т. Т.* 3. Москва, Мысль, 1984, с. 494–500.

[9] Лейбниц Г.В. *Письмо к герцогу Ганноверскому. Сочинения в 4 т. Т.* 3. Москва, Мысль, 1984, с. 491–493.

[10] Лейбниц Г.В. *Письма и эссе о китайской философии и двоичной системе исчисления*. Москва, Институт философии РАН, 2005, 404 с.

[11] Ершов Ю.А., Целищев В.В. *Алгоритмы и вычислимость в человеческом познании*. Новосибирск, СО РАН, 2012, 505 с.

[12] Лобовиков В.О. *«Искусственный интеллект», формальная этика и морально-правовой выбор*. Свердловск, Уральский университет, 1988, 188 с.

[13] Лобовиков В.О. *Математическая этика, метафизика и естественное право (Алгебра метафизики как алгебра формальной аксиологии)*. Екатеринбург, Институт философии и права Уральского отделения РАН, 2007, 408 с.

[14] Lobovikov V. Algebra of Morality and Formal Ethics. Looking Back to See the Future: Reflections on Sins and Virtues. Oxford, United Kingdom, Inter-Disciplinary Press, 2014, pp. 17–41.

[15] Лобовиков В.О. *Алгебра формальной аксиологии как дискретная математическая модель философии: основания для конструирования метафизической подсистемы искусственного интеллекта автономного когнитивного робота*. Москва, Академический проект, 2019, с. 244–285.

[16] Lobovikov V.O. Artificial Intelligence and an almost Unknown Aspect of Mathematical Linguistics. Discourse-P, 2022, vol. 19, no. 3, pp. 170–184. https://doi.org/10.17506/18179568_2022_19_3_170

[17] Мальцев А.И. *Алгебраические системы*. Москва, Наука, 1970, 392 с.