А. В. РАЗИН

ЗАМЕТКИ ПО ПОВОДУ ЗАКОНОВ РОБОТОТЕХНИКИ А. АЗИМОВА*

В статье рассматриваются вопросы оснований человеческой деятельности, предполагающей наличие у людей сознания и самосознания, средств коммуникации, основанных на языке, организованных форм познания и культурной принадлежности индивида. Показывается, что в отличие от искусственного интеллекта человек способен к поиску принципиально новых средств достижения целей, а сами цели выглядят как утверждаемые на основе динамического процесса развития культуры всего общества и вовлеченного в этот процесс индивида. В то же время, если робот действует по приказам человека, он фактически в собственном целеполагании не нуждается. Согласно авторской позиции законы Азимова не учитывают задач, которые в принципе способен осуществлять искусственный интеллект в современном обществе, и потому они не могут быть использованы в их непосредственном виде, а должны быть заменены дифференцированными кодексами, на которые смогут опираться разработчики робототехники.

Ключевые слова: *робототехника, законы, сознание, самосознание, программа, творчество, этика, мораль, дискурс, легитимные решения.*

The article examines the foundations of human activity, which presupposes the presence of consciousness and self-awareness as a fundamental human character, means of communication based on language, organized forms of cognition and cultural affiliation of the individual. It is shown that a human being, unlike an artificial intelligence, is capable of searching for fundamentally new means of achieving goals, while the goals themselves are created through the dynamic process of cultural development of society as a whole and the involvement of the individual in this process. At the same time, if a robot

* ***Для цитирования:*** Разин А. В. Заметки по поводу законов робототехники А. Азимова // Философия и общество. 2024. № 4. С. 37–48. DOI: 10.30884/jfio/2024.04.03.

For citation: Razin A. V. Notes on Isaac Asimov's Laws of Robotics // Filosofiya i obshchestvo = Philosophy and Society. 2024. No. 4. Pp. 37–48. DOI: 10.30884/jfio/2024.04.03 (in Russian).

acts on the basis of human orders, it does not really need to set its own goals. According to the author, Asimov's laws do not take into account the tasks that artificial intelligence is in principle capable of performing in modern society, and therefore they cannot be used in their direct form, but should be replaced by differentiated codes that robotics developers can rely on.

Keywords: *robotics, laws, consciousness, self-awareness, program, creativity, ethics, morality, discourse, legitimate decisions.*

Введение

А. Азимов – известный писатель-фантаст, по образованию биохимик, ученый и популяризатор науки – сформулировал три закона робототехники, которые, как он полагал, могут позволить предотвратить возможный вред от использования роботов человеком и в то же время превратить их в эффективных помощников людей. Но в настоящее время меняются и сами представления о роботах, и представления о тех задачах, которые можно на роботов возложить. Кроме того, является актуальным и вопрос о том, насколько робот может реально обладать сознанием и самосознанием, насколько он способен выйти за пределы заложенной в него программы действия, оказаться способным к творчеству и стать реальным моральным агентом. Это заставляет вернуться к проблеме законов робототехники. В данной статье мы, во-первых, рассмотрим общие условия человеческой деятельности, предполагающей сознание и самосознание, во-вторых, в основном будем опираться на материалы дискуссии между Беном Герцелем (Aidyia Holdings) и Луисом Хельмом, заместителем директора Института исследований машинного интеллекта, описанные в интервью Ильи Хеля под названием «Смогут ли "три закона робототехники" защитить нас?».

Основания деятельности

В основании деятельности человека лежат его потребности и осознание условий их удовлетворения, в том числе – осознание и исполнение обязательств перед другими людьми. Таким образом, человек действует на основании собственных желаний и представлений о средствах их удовлетворения. Соединение этих двух моментов позволяет сформулировать цели деятельности. Робот, как это следует из второго закона Азимова, действует по приказу человека. Следовательно, собственных целей бытия он фактически не имеет.

Насколько при этом робот может обладать мышлением, осознавать сам себя и, наконец, быть моральным агентом, представляется большим вопросом. Достаточно просто создать у робота некоторую имитацию эмоциональной жизни: цель, заданная человеком, оценка способа ее достижения на основе вариантов действий, заложенных в программу (в данные, которые могут быть очень большими, но в принципе уже известными в человеческой практике), возбуждение в связи с близким достижением цели (возрастание ресурсов), разочарование при недостижении цели (переход к экономии ресурсов), в конечном счете – отказ от действия. Но это лишь похоже на эмоциональную жизнь человека.

Начнем с того, что такое потребность. Это осознанная нужда, например сознание недостатка чего-либо, необходимого для поддержания жизни человека. Могут быть нужды (например, нехватка витаминов), которые не осознаются человеком. Для их осознания нужны специальные процедуры – скажем, медицинская помощь.

Стоп: потребность осознанная нужна, следовательно, для того чтобы иметь потребность, нужно иметь сознание и самосознание. А возможно ли оно у робота, да и нужно ли оно ему, если он действует не на основе потребностей, а на основе приказов, которые ему отдает человек?

В ряде статей мы показали множество условий, необходимых для того, чтобы у искусственного интеллекта можно было сформировать сознание: феноменальный опыт, тело, состояние которого постоянно контролируется и оценивается эмоционально, язык и связанная с ним способность к абстрактному мышлению, самостоятельное предметное освоение мира (незаинтересованное, не связанное с конкретным действием познание реальности), общение с себе подобными, значимость их оценок, классификация собственных желаний на более и менее значимые, так же как и оценка того, какие желания надо удовлетворять в первую очередь и надо ли их вообще удовлетворять или лучше от них оказаться [Разин 2019; 2023]. Как очень точно заметил М. Шелер, человек единственный, кто может сказать миру нет. «...Человек есть то живое существо, которое может (подавляя и вытесняя импульсы собственных влечений, отказывая им в питании образами восприятия и представлениями) относиться принципиально аскетически к своей жизни, вселяющей в него ужас. По сравнению с животным, ко-

торое всегда говорит “да” действительному бытию, даже если пугается и бежит, человек – это “тот, кто может сказать нет”, “аскет жизни”, вечный протестант против всякой только действительности» [Шелер 1988: 65]. И этим отрицанием, способностью отказа от одних желаний ради других, собственно, обусловлено все духовное, в том числе и моральное развитие человека.

«У животного – высоко- или низкоорганизованного – всякое действие, всякая реакция, которую оно производит, в том числе и “разумная”, исходят из физиологической определенности его нервной системы, которой в области психики подчинены импульсы влечений и чувственное восприятие. Что не интересно для этих влечений, то и не дано, а что дано, то дано лишь как центр сопротивления его желанию и отвращению» [Там же: 54]. Человек же способен к анализу собственного бытия с точки зрения поиска предпочтения одних ценностей другим, к отказу от одних желаний ради других, которые рассматриваются как более высокие, в большей мере отвечающие представлениям об идеалах жизни, о подлинной человеческой сущности.

А способен ли на подобное отрицание, на предпочтение одних ценностей другим робот? Если он только выполняет распоряжения человека, то ему думать о предпочтениях и не надо. Но, может быть, когда-то появятся такие интегрированные в человеческие сообщества роботы (андроиды), у которых будут свои цели, и они станут выполнять распоряжения других людей (обладающих властными полномочиями) не более, чем мы с вами?

Вопрос о том, насколько робот, даже андроид, обладающий схожим с человеком телом и усвоивший какие-то области человеческой культуры, может быть реально интегрирован в человеческое сообщество, представляется весьма сложным, и ответ на него, скорее всего, будет отрицательным. Дело здесь в субъективности, которая зависит от феноменального опыта и от квалиативных состояний сознания. Эти состояния зависят от многих причин, но в немалой степени определяются тем, что нейроны головного мозга представляют собой живые клетки, которые могут быть специализированными, имеют свои цели, объединяются с другими нейронами по еще не известным нам до конца принципам.

В настоящее время ясно, что сознание имеет гипотезотворческую и даже галлюциногенную природу. В мозге есть петлевые

процессы (Дж. Ризолатти, Н. Хамфри), то есть он постоянно играет сам с собой, и это определяет его творческие возможности. Такое творчество в конечном счете способствует тому, что могут быть созданы совершенно новые средства достижения целей. А если у робота нет сознания, он и не способен на создание принципиально нового.

Мозг человека, конечно, имеет некоторые алгоритмы действия. Человек знает, что такое правильные рассуждения, какие бывают логические операции, но это далеко не все его способности. Логика вообще не способна произвести принципиально новое знание, в выводах не сможет быть того, чего нет в посылках. Поэтому с помощью логики можно только преобразовать уже имеющееся знание. Но разве человеческий процесс познания сводится только к этому? Он предполагает создание гипотез, выдвижение предположений о том, как устроен мир, при этом человек постоянно стремится к расширению горизонта познания, формулированию таких утверждений, которые позволят найти новые основания для предполагаемых утверждений. Например, бессмысленно говорить, что́ вокруг чего вращается, Земля вокруг Солнца или Солнце со всем небесным сводом вокруг Земли. Для описания реальности может быть использована и та и другая схема. А для окончательного утверждения о том, что именно Земля вращается вокруг Солнца, надо сформулировать закон всемирного тяготения.

Способен ли искусственный интеллект к такому расширению ви́дения мира? Думается, что нет, и он не приобретет такую способность, пока у него не будет развитого сознания и пока он не будет интегрирован в систему научных представлений и методов приобретения научного знания, что само по себе представляет специальную, динамически развивающуюся область культуры. Соответственно, для такой интеграции надо не только иметь логику мышления, сходную с логикой мышления человека, но и постоянно осваивать новые области культуры, причем культура в данном случае совсем не просто некоторая база данных о достижениях человечества в каких-то областях, а именно динамика общественного и личного бытия.

Современный же искусственный интеллект даже не оперирует понятием «переменная», то есть и логика, по которой он принимает решения, весьма скудна, хотя работать он может очень быстро.

Приступая к выполнению какого-то задания, робот в принципе работает по линейной схеме. Цель, заданная распоряжениями человека, обращение к базам данных, находящихся в разных информационных источниках, сравнение данных с точки зрения выбора оптимального пути. Попытка совершения действия, направленного на достижение цели, которое может быть удачным или неудачным. В случае удачи можно сделать так, что загораются лампочки красного цвета, а в случае неудачи – черного цвета. Это будет внешний сигнал выражения какого-то аналога радости и разочарования, доступный для восприятия со стороны других интеллектуальных систем, что важно, если предположить трансляцию удачного опыта в некотором машинном сообществе, владеющего информацией о совместных целях деятельности (пока такого сообщества, конечно, нет). При удаче может происходить самообучение, то есть удачное действие может запоминаться, в случае неудачи – запоминаться, что так действовать нельзя. Но цвет лампочек – это в данном случае просто внешний сигнал, ничего общего с человеческими эмоциями не имеющий. А почему? А потому, что эмоции связаны с особого типа удовольствием или неудовольствием – разочарованием. И особенность человеческой деятельности заключается в том, что человек может начинать действовать ради тех эмоций, которые он когда-то испытал, они могут стать новыми самостоятельными ценностями и привести к формированию новых потребностей. «...Человеческие эмоции способны одновременно выступать в двух разных социальных ролях: в роли оценок, помечающих предмет деятельности и регулирующих ее ход, и в роли самодовлеющих ценностей, которые обогащают и превращают в дополнительный мотив деятельности сам ее процесс» [Додонов 1978: 244]. Но у искусственного интеллекта, по крайней мере на данный момент, нет такой субъективности, которая может превратить эмоцию в самостоятельную ценность.

Далее, как уже было сказано, искусственный интеллект в решении задачи действует по линейной схеме. Он не способен к расширению масштабов своей деятельности. Когда распоряжение человека выполнено, действие завершается. Но сознание человека не работает по линейным схемам. В нем нет единого управляющего центра, его особенность скорее выражается принципами сетевых связей, причем для решения конкретной задачи могут создаваться

новые цепи, подключаться новые нейроны. В какой-то степени это пытаются моделировать в нейросетях, но они далеки от совершенства, так как по существу работают просто на основе сравнения данных, в них нет семантики, нет субъективной заинтересованности в жизни, как у живых организмов. Очень распространено мнение, что нейросети способны к самообучению, к решению творческих задач. Но в действительности любая нейросеть первоначально настраивается программистом, а ее творческий поиск задается так называемым промптом, то есть вопросами (подсказками), которые задаются тем, кто работает с нейросетью.

Можно создать у робота какой-то сигнал, который внешне будет напоминать некоторую фиксированную нужду, но отнюдь не потребность, так как потребность, как уже было сказано, это осознанная нужда, и для того, чтобы она выступала в качестве основания начала деятельности, необходимо сознание.

Далее необходимо обратить внимание на условия удовлетворения потребностей. Дж. Серль прекрасно показал, что совсем не все, что делает человек, обусловлено его потребностями. Например, если я хочу утолить жажду с помощью пива, я должен (нормативное требование) за него заплатить, и сама такая плата потребностью не является. Это условие социального бытия человека. «…В человеческой рациональности, в противоположность обезьяньей, существует разница между основаниями для действия, которые связаны с удовлетворением того или иного желания, и основаниями, не зависящими от желаний. Основное различие между видами оснований для действия заключается в том, что одни из них связаны с тем, что вы хотите и что вы должны сделать для получения желаемого, тогда как другие связаны с тем, что вы должны сделать независимо от ваших желаний» [Серль 2004: 21].

Какие-то условия социального бытия в искусственный интеллект можно заложить. Но вот беда, у человека есть свобода выбора, связанная с его субъективностью и субъективной оценкой всех условий жизни. Например, если нет денег, можно не покупать пиво, а выпить воды. Рациональность поведения человека касается не только рациональности средств, но и рациональности целей. Человек сталкивается с проблемой предпочтения одних целей другим, оценки вероятности их реализации, рисков реализации. Более того, вся деятельность человека развивается в горизонте осознания своей

конечности. С этим связываются жизненные планы, предпочтения, задается темп жизни. Все это вряд ли будет актуальным для робота.

Законы А. Азимова

Но вернемся к законам А. Азимова. Один закон мы уже упомянули – робот должен выполнять распоряжения человека. Это второй закон. Таким образом робот уже лишается самостоятельности и претензий на самостоятельную жизнь.

В последовательном изложении все предложенные Азимовым законы выглядят следующим образом:

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред.

2. Робот должен повиноваться всем приказам, которые дает человек, кроме тех случаев, когда эти приказы противоречат Первому закону.

3. Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит Первому и Второму законам.

0. Робот не может причинить вреда человеку, если только он не докажет, что в конечном счете это будет полезно для всего человечества.

Иначе: «Робот не может причинять вред человечеству или своим бездействием допустить, чтобы человечеству был причинен вред».

Наибольшие вопросы вызывает именно этот последний закон. Что значит своим бездействием причинить вред человечеству? А что если для предотвращения вреда надо кого-то убить, например ученого, который ставит опасные эксперименты или сделал такое открытие, которое создаст человечеству огромные проблемы, скажем, изобрел средство радикального продления жизни, или если ради борьбы с опасной инфекцией надо уничтожить целый город? Здесь возникает очень непростая проблема приоритетов, вопрос о том, какую этическую концепцию мы должны принять, можем ли мы допустить выбор в сторону меньшего зла, рассуждая с позиции утилитаризма, или же мы должны исходить из деонтологии, позиции абсолютного понимания морали, в соответствии с которой недопустимо использовать человека в качестве средства?

Другие законы тоже вызывают вопросы, например:

• О какой временной перспективе идет речь? Как это связано с принципом разумной предосторожности, нанесения вреда экологии?

• Как это соответствует условиям свободной конкуренции? Не распространяя какое-то ноу-хау, я в общем-то причиняю вред человечеству?

• Обладает ли робот способностью оценить вред?

Мы полагаем, что одна из основных проблем, связанная с распоряжениями человека, отданными роботу, заключается в вопросе о легитимности данных распоряжений. Конечно, если робот выполняет поручения типа принести из дома забытую сумочку, азимовские законы в своей базовой части вполне подходят. А если речь идет о производственных или военных решениях? Допустим, группа людей, политиков считает, что некоторой стране надо передать или продать какое-то вооружение. Не обязательно предполагается, что оно будет использовано для прямого вреда человеку. Может быть, просто как фактор сдерживания возможной агрессии другого государства. Но некоторые предполагают, что подобный акт представляет прямую угрозу применения такого оружия. Робот вполне может быть вовлечен в процесс передачи вооружения, и как он в таком случае должен поступить: саботировать такой процесс, или же нет? То же относится и к технологиям, которые могут рассматриваться как опасные. Ряд стран использует атомную энергетику, некоторые решили отказаться от этого после аварии на «Фукусиме». Кого роботы должны слушать в такой ситуации? Ясно, что сами распоряжения, которые будет выполнять искусственный интеллект, предполагают предварительное обсуждение легитимности принимаемых решений. Это предполагает процедуру дискурса, который, конечно, может быть ограничен рамками национальных государств и решением их правительств, но уже сейчас ясно, что такой дискурс преодолевает национальные границы. Или другой пример: обсуждается вопрос о том, можно ли строить аэродром на нестабильных почвах, скажем, в условиях вечной мерзлоты. Ясно, что без создания сложных технических систем, направленных на локализацию возможного вреда экологии, строительство принесет очевидный вред природе. Но создание подобных систем обойдется дорого в экономическом смысле. Высказывается другое мнение: построить временные взлетно-посадочные полосы. Это даст возможность перебросить грузы, развить регион, создать там промышленность и т. д. Конечно, природе будет нанесен вред, но новые индустриальные возможности, новые технологии позволят его устра-

нить в будущем. Понятно, что дискуссия должна быть как-то завершена, прежде чем роботы будут вовлечены в исполнение решений, и не самим роботам судить о том, какое решение правильное.

Разработчики Этического руководства для программирования автоматизированных систем вождения (Германия, 2017 г.) сами говорят, что их правила основаны на деонтологии. Но беда деонтологии – в предельном ограничении свободы выбора, отказе от ситуативных решений. Принципы выстраиваются в строго лексической последовательности. Одни однозначно подчиняются другим (так же, кстати, построены и законы А. Азимова). Многие считают деонтологию недостаточной этической теорией. Утилитаризм более гибок, он предполагает поиск компромисса между разными принципами.

Специалисты обсуждают вопрос о том, в каких формах вообще будет существовать искусственный интеллект. Так, бразилец Бен Герцель, главный научный сотрудник компании Aidyia Holdings, отмечает: «Думаю, что тип роботов, которых предвидел Азимов, станет возможным в недалеком будущем. Тем не менее в большинстве своих вымышленных миров писатель предполагал, что человекоподобные роботы будут вершиной робототехники и инженерии искусственного интеллекта. Это вряд ли. Очень скоро, после достижения статуса азимовских роботов, станет доступным и создание искусственного сверхинтеллекта и сверхроботов» [Хель 2014]. Б. Герцель считает, что искусственный суперинтеллект будет существовать в сетях, а сам человек будет киборгизироваться.

Мы подвергаем сомнению существование роботов, подобных азимовским, во всяком случае, для этого надо выполнить очень многие условия, и это весьма отдаленная перспектива. А вот существование развитого искусственного интеллекта в сетях кажется весьма вероятным.

Что же касается самих азимовских законов, то их недостатки, как считает Луис Хельм, видятся в следующем:

- Они состязательны по своей сути;
- основаны на изжившей себя этической теории (деонтологии);
- не работают даже в фантастике [Там же].

Б. Герцель соглашается с общей оценкой законов: в реальности эти законы работать не будут, поскольку термины с их участием неоднозначны и остаются предметом толкования – а значит, крайне зависимы от тех, кто делает переводы [Хель 2014].

Л. Хель считает, что создавать робота с самосознанием или отдельными его элементами, позволяющими осуществлять самостоятельный выбор, вообще не нужно. Но тогда совершенно ясно, что вся ответственность за принятые искусственным интеллектом решения ляжет на разработчиков и тех, кто определяет сферу его применения.

Азимовские законы также упрекают в шовинизме по отношению к роботам, ставится вопрос о том, почему искусственный интеллект должен быть всецело подчинен человеку, если он постепенно набирает все большую мощь. Здесь ставятся вопросы о том, морально ли будет выключать роботы, использовать их в качестве рабов для выполнения тяжелой работы, хотя исходно в пьесе К. Чапека «Россумские универсальные роботы», в которой и было введено понятие «робот», как раз и предполагалось, что роботы создаются для выполнения тяжелой работы (хотя в конце концов они против этого восстали).

Заключение

В настоящей статье мы пытались показать принципиальные отличия условий деятельности искусственного интеллекта, роботов-андроидов (если/и когда они будут созданы) от условий и оснований деятельности человека. Принципиально то, что сознание человека формируется в связи с его феноменальным опытом и оценкой общей перспективы жизни. Робот по определению рассматривается как в принципе бессмертное существо. Так что вряд ли он даже в перспективе будет обладать сознанием, схожим с сознанием человека. В данной связи анализируются и законы Азимова. В них прежде всего не учитывается, в решение каких задач будут вовлечены роботы, кто будет отвечать за определение самих этих задач, а если исходить из того, что робот создается, чтобы выполнять приказы человека, то вопрос об их возможном сознании и самосознании вообще снимается. Остается также непонятным, почему робот может и/или должен думать о собственной безопасности, если у него нет идеи ценности актуального бытия.

И последнее. Если мы пойдем по пути полной имитации человека техническими системами, создадим аналоги мозговых нейронов, дадим роботу находиться в мире иллюзий, ограничим период его жизни, чтобы он смог определять приоритетные цели соб-

ственного бытия, не потеряем ли мы все те преимущества, которые дает нам искусственный интеллект, то есть быстроту решений, действия по строгому алгоритму, недопущение ошибок и т. д.?

Литература

Додонов Б. И. Эмоция как ценность. М. : Политиздат, 1978.

Разин А. В. Этика искусственного интеллекта // Философия и общество. 2019. № 1. С. 57–73.

Разин А. В. Компьютер и мозг: проблема квалиа // Философия и общество. 2023. № 1. С. 42–56.

Серль Дж. Рациональность в действии. М. : Прогресс-Традиция, 2004.

Хель И. Смогут ли три закона робототехники защитить нас? 2014 [Электронный ресурс]. URL: https://hi-news.ru/robots/smogut-li-tri-zakona-robototexniki-zashhitit-nas.html (дата обращения: 16.09.2024).

Шелер М. Положение человека в космосе // Проблема человека в западной философии. М. : Прогресс, 1988. С. 31–95.