В. Э. КАРПОВ

К ВОПРОСУ О ЗАКОНАХ РОБОТОТЕХНИКИ А. АЗИМОВА*

Работа посвящена обсуждению некоторых аспектов известных законов робототехники А. Азимова. Раскрыто утверждение о том, что рассуждения о противоречиях законов, их критика не являются конструктивными и что сам автор этих законов понимал существующие коллизии формулировок крайне глубоко. В работе показано, что поведение азимовских роботов определяется не столько этими законами, сколько особенностями их эмоционально-потребностной архитектуры. При этом сами законы тесно связаны с вопросами этики автономных/интеллектуальных систем. Кроме того, азимовские законы вскрывают ряд сугубо технических проблем, которые не решены до сих пор. В частности, нерешенным остается вопрос о том, на каком уровне должны быть реализованы этические императивы поведения робота.

Ключевые слова: *законы робототехники, когнитивная система управления, эмоционально-потребностная архитектура, этика, коллаборативный робот.*

The paper is devoted to the discussion of some aspects of the well-known laws of robotics by Isaac Asimov. The author reveals the statement that reasoning about the contradictions of laws and their criticism ise not constructive and that IsaacAsimov himself understood the existing conflicts of formulations very deeply. The paper shows that the behavior of Asimov's robots is determined less by these laws than by the peculiarities of architecture of their emotions and needs. At the same time, the laws themselves are closely related to the ethical issues of autonomous/intelligent systems. In addition, Asimov's laws reveal

* Работа проведена в рамках выполнения государственного задания НИЦ «Курчатовский институт».

Для цитирования: Карпов В. Э. К вопросу о законах робототехники А. Азимова // Философия и общество. 2024. № 4. С. 19–36. DOI: 10.30884/jfio/2024.04.02.

For citation: Karpov V. E. On the Laws of Robotics by Isaac Asimov // Filosofiya i obshchestvo = Philosophy and Society. 2024. No. 4. Pp. 19–36. DOI: 10.30884/jfio/2024.04.02 (in Russian).

a number of purely technical problems that remain unsolved. In particular, the question of the level at which the ethical imperatives of robot behavior should be implemented remains unresolved.

Keywords: *laws of robotics, cognitive control system, emotions-needs architecture, ethics, collaborative robot.*

1. Введение

К сожалению, к статьям не пишут предисловий. В противном случае настоящий текст начинался бы с рассуждений о том, что сегодня сложилась парадоксальная и противоестественная ситуация, когда обсуждение специальных, сугубо научных и технических вопросов стало крайне модным в непрофессиональной, гуманитарной среде. Причем популярность темы того же искусственного интеллекта, активность гуманитарного сообщества зачастую приводят к негативным последствиям для сообщества профессионального: принимаются странные законы, переориентируются акценты, общество в целом получает совершенно искаженную картину состояния и перспектив развития науки и техники. Эту противоестественность хорошо иллюстрирует такой казус, как восторженное восприятие гуманитарным сообществом явных симулякров. Так, общество понимает, что ему преподносится некая имитация осмысленности в виде, например, больших языковых моделей (всевозможные чаты GPT) или хорошо натасканных на каких-то примерах нейронных сетей. Но вместо критического и скептического отношения к подобного рода вещам разворачиваются обсуждения, дискуссии, принимаются административные решения, строятся планы по использованию таких механизмов в госуправлении, юриспруденции и т. п.

В романтический период становления ИИ картина была диаметрально противоположной. Общество – гуманитарная часть – активно возмущалось, когда вместо чего-то реально подобного человеческому разуму ему предлагались внешние имитации, игры в осмысленность – см., например, классические книги Х. Дрейфуса [1978] или Дж. Вейценбаума [1982]. Эти объективно актуальные и глубокие труды сегодня уже не воспринимаются публикой. Но помимо нейронных сетей есть еще одна тема, «оседланная» гуманитарным сообществом. Это пресловутые законы робототехники

А. Азимова. Интересно, что в профессиональном робототехническом сообществе данная тема почти не затрагивается, и совершенно напрасно. Здесь есть о чем подумать, здесь присутствует множество междисциплинарных проблем. Исходя из всего этого и возникла необходимость обсуждения законов Азимова.

Проблема лишь в том, что предлагаемый ниже текст не имеет определенного адресата. Очевидны будут возражения и критические замечания философов. Многое из сказанного было ими давно обсуждено, изучено, разложено и проанализировано. Будут, естественно, замечания и контрдоводы со стороны представителей строгих, технических наук. Расширять изложение, превращать ограниченный текст в обширный труд, строгий, полный, с массой ссылок и разъяснений – это уже другая задача. Здесь лишь хочется заметить, что основные положения статьи, касающиеся «технических» аспектов, основаны на реальных проектах и исследованиях. И если автор рассуждает об эмоциональной архитектуре робота, реализации механизмов эмпатии, социуме роботов, то это тоже обоснованно хотя бы с точки зрения его – автора – профессионального рода деятельности.

Итак, вернемся к теме статьи, полагая, что ее цель – обозначить некоторые аспекты азимовских законов, на которые далеко не всегда обращают внимание или вовсе вкладывают в них иной смысл.

Законы робототехники Айзека Азимова – это, пожалуй, одна из самых обсуждаемых тем в изучении взаимоотношений автономных/интеллектуальных систем и человека. Начиная с 1940-х гг. [Asimov 1940] эти законы неоднократно анализировались и дополнялись. Впервые в явном виде они появляются в рассказе Айзека Азимова «Хоровод» в марте 1942 г. (менее явно законы фигурируют в рассказе «Лжец!», написанном в 1941 г.). Сами законы выглядят так [Азимов 2022]:

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

2. Робот должен повиноваться командам человека, если эти команды не противоречат Первому закону.

3. Робот должен заботиться о своей безопасности, поскольку [в той мере, если] это не противоречит Первому и Второму законам.

Примечание. В рассказе «Хоровод» описывается ситуация конфликта между стремлением робота заботиться о своей безопасности (Третий закон) и необходимостью подчинения (Второй закон). Роботу было дано задание отправиться за крайне необходимым героям рассказа селеном, однако в районе этого селенового озера была неблагоприятная для робота обстановка. В итоге у робота по мере приближения к опасному участку начинал преобладать императив самосохранения, а при удалении от него начинал превалировать закон повиновения. Начинался «хоровод»: «Третий закон гонит его назад, а Второй – вперед... И он начинает кружить около озера, оставаясь на линии, где существует это равновесие».

Далее мы обратим внимание на некоторые не столь очевидные аспекты законов Азимова и покажем, что:

1. Азимов писал как раз о противоречиях этих законов.

2. То, о чем писал Азимов в его «Хороводе», – это вовсе не о законах робототехники, а об эмоциях.

3. В этих законах имеется ряд сугубо технических (робототехнических) проблем, не решенных до сих пор.

4. Законы Азимова имеют тесную связь с вопросами этики автономных/интеллектуальных систем.

2. Внешние и внутренние противоречия

Под внешними противоречиями формулировок законов мы будем понимать то, что лежит на поверхности, – их форму, которая активно и широко обсуждается и к которой сам их автор относился весьма критически. А внутренние противоречия законов – это про подмену понятий.

2.1. Внешние противоречия

Подобного рода формулировки очень популярны в гуманитарной среде в силу своей понятности и очевидности. Более того, перечень законов активно пополнялся и уточнялся. Так, в 1986 г. Азимов предложил так называемый нулевой закон с максимальным приоритетом: «Робот не может причинить вред человечеству или своим бездействием допустить, чтобы человечеству был причинен вред». Более того, пробуждение активного интереса гуманитарной общественности к вопросам этики систем искусственного интел-

лекта (более строго – автономных/интеллектуальных систем – АИС) привело к тому, что в различного рода обсуждениях и публикациях по этим вопросам упоминание или ссылка на эти три закона является почти обязательным. Мы не будем приводить ссылки на эти многочисленные источники. Даже самый поверхностный запрос в РИНЦ выдает список почти из двух десятков публикаций, в которых упоминаются эти законы (в основном это публикации из области философии, антропологии и даже юриспруденции).

Но при этом обычно забывают, что А. Азимов – не только писатель-фантаст, но и ученый (биохимик), популяризатор и историк науки. Он понимал, что пишет, что и как формулирует. Смысл как раз в том, что его рассказы про роботов – истории коллизий этих законов. В такие «очевидные», понятные и всех устраивающие формулировки заложены сплошные противоречия, неопределенности, конфликты, которые и определяют фабулу и драматизм историй Азимова. Иными словами, внешне законы выглядят хорошо, но в таких формулировках они работать не могут. Однако об этом будет сказано ниже.

2.2. Эмоции и поведение

На самом деле в упомянутом рассказе А. Азимова «Хоровод» поведение робота имеет весьма косвенное отношение к трем законам. Причиной неустойчивого поведения робота являлась его эмоциональная несбалансированность. Эмоции как комплекс процессов присущи не только человеку или высшим млекопитающим. Эмоции – это нижний, психофизиологический уровень управления животного, обладающего более или менее сложным поведением. При этом основой эмоциональных процессов является интегральная оценка ситуации, когда определяется баланс между теми средствами и ресурсами, которые нужны животному для удовлетворения своих актуальных потребностей, и теми, которые имеются в наличии. Если баланс положителен, то речь идет о положительных эмоциях, в противном случае – об отрицательных. Или, иначе, согласно В. П. Симонову, эмоции являются оценкой текущей потребности (ее качества и ценности) и возможности ее удовлетворения [Симонов 1982; Simonov 1991].

В общем виде отношение этих факторов описывается качественным (оценочным) выражением:

$$E = f(N, p(I_{need}, I_{has})), \quad (1)$$

где E – эмоция, ее величина и знак (качество); N – сила и качество текущей необходимости; $p(I_{need}, I_{has})$ – оценка возможности удовлетворить потребность на базе врожденного и полученного жизненного опыта; I_{need} – информация о способе удовлетворения потребности; I_{has} – информация об имеющихся у субъекта средствах, ресурсах и времени. Мы можем объяснить выражение (1) следующим образом: индивид оценивает свои текущие потребности I_{need} или то, что он должен сделать (поесть, найти еду, отдохнуть, убежать от опасности и т. д.). Затем он оценивает индивидуальные возможности удовлетворения этих потребностей I_{has}. Различие между потребностями и возможностями определяет эмоциональную оценку текущей ситуации. Если он имеет некоторые потребности и при этом возможности для их удовлетворения достаточны, индивид получает положительную эмоциональную оценку. В противном случае его эмоции негативны.

Именно эмоции отвечают за стабилизацию поведения, контрастирование восприятия, даже за кратковременную память.

Если говорить о реализации механизма эмоций, то речь идет о контурах обратной связи в системе управления, которые отвечают за устойчивость поведения. Нарушение эмоционального механизма влечет серьезные функциональные нарушения, известные в нейропсихологии, подобные тем, что были у робота Спиди из «Хоровода». Подробнее об эмоциях и о темпераменте робота см.: [Карпов 2014; Karpov 2014].

На Рис. 1 приведен фрагмент эмоционально-потребностной схемы системы управления робота Спиди. Здесь мы предполагаем, что одной из базовых потребностей робота является потребность в подчинении человеку, а целевой объект (селен, за которым посылали робота) обозначен как «Цель». Самым интересным элементом схемы является шлюз – тот узел, который и формирует процессы, называемые эмоциями. С каждым шлюзом связана своя частная эмоция, которая зависит от того, какова сила соответствующей потребности, что наблюдает робот (сенсорика) и чем он реально занят. Именно шлюз ответствен за сенсорику и фантомные сигналы

(раздражителя уже нет, опасность не видна, а агент продолжает убегать), именно там накапливается отрицательная эмоция, связанная с тем, что агент не выполняет сейчас ту поведенческую процедуру, которая должна быть актуальной и т. д.

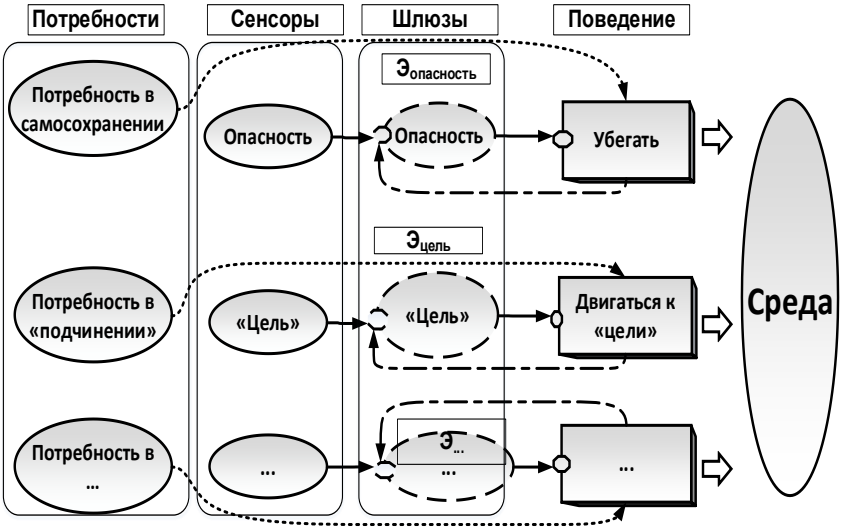

Рис. 1. Фрагмент эмоционально-потребностной схемы поведения робота Спиди

Иными словами, поведенческая процедура, которую выполняет робот, зависит напрямую от эмоций, определяемых степенью удовлетворения потребностей. Роботу отдан приказ доставить селен, потребность в этом актуальна и велика. Более того, известно, куда двигаться и что с ним делать, однако вместо этого робот вынужден удаляться от опасного места (потребность в самосохранении). Через какое-то время накапливается отрицательная эмоция, формируемая шлюзом «Цель». К тому же ослабевает активность процедуры убегания (робот удаляется от опасности). В результате робот начнет выполнять процедуру движения к цели, приближаясь при этом к опасному участку. И все повторяется.

На самом деле такое поведение можно «отрегулировать», изменяя силу связей между элементами схемы, то есть сделать его более эмоционально устойчивым. Кстати, «подкручивая» параметры этой схемы, мы получим различные формы поведения робота. При

одних параметрах робот будет быстро и активно реагировать на малейшие изменения внешних сигналов, при других, напротив, – медленно переключаться между поведенческими процедурами и т. д. Такие типы реакций поведения в целом удобно описывать в терминах темперамента. Спиди, судя по всему, был типичным холериком.

Напоследок важно отметить еще и то, что зачастую эмоции путают с чувствами, а сами эмоциональные процессы – с внешним проявлением эмоций. И то и другое в корне неверно. Чувства направлены на что-то конкретное, а эмоции – интегральная оценка ситуации, при этом репертуар эмоций крайне скуден, что бы ни утверждали последователи Р. Плутчика [Plutchik 2001] и им подобные, рисуя красивые картинки и безудержно расширяя спектр эмоций. Согласно работе [Симонов 1987], есть лишь четыре эмоциональных состояния: гнев, страх, удовольствие и отвращение. И все они связаны с базовыми поведенческими процедурами.

3. Технический аспект

В законах робототехники есть крайне интересный и важный технический аспект, на который обычно не обращают особого внимания. Фактически автор писал о научных, технических проблемах, не решенных до сих пор. По мнению А. Азимова, структура мозга робота (системы управления) должна быть такой, чтобы эта система вообще не могла функционировать при нарушении определенных правил поведения (законов робототехники). Это означает, что законы не просто органично интегрированы в структуру системы управления (СУ), но образуют центральную ось всей структуры СУ. В этом смысле можно поставить вопросы: возможно ли создать такую архитектуру системы управления АИС, чтобы в ее основе лежали этические императивы, а не просто какой-то дополнительный набор эвристик? Чем будет руководствоваться АИС при принятии решений? Или, как вариант, могут ли этические императивы стать важной частью каких-либо потребностей АИС (наряду с потребностью самосохранения, реализации социальных функций и т. д.)?

Рассмотрим далее некоторые сугубо «технические» аспекты, связанные с азимовскими законами робототехники.

3.1. Система управления роботом и этика поведения

Для начала отметим, что азимовские законы имеют теснейшую связь с этикой. В рассказе «Улика» того же сборника «Я, робот» Азимов пишет: «…Три Закона Роботехники совпадают с основными принципами большинства этических систем, существующих на Земле». Далее синонимичные понятия «этика», «мораль», «нравственность» будут встречаться часто, поэтому отметим сразу два важных аспекта.

1. Основное назначение этики – разрешение конфликтов [Гусейнов, Апресян 2000]. Мораль – это адаптивный механизм, позволяющий социуму более эффективно приспосабливаться к сложным условиям.

2. Мораль – это вовсе не прерогатива человека [де Вааль 2019]. Например, такое базовое понятие этического поведения, как эмпатия, – это сугубо биологический механизм, присущий животным, причем не только высшим млекопитающим. То же самое относится к понятиям «Я», «свой», «чужой», которые определяют социальное поведение животного и являются основой того, что называется «золотым правилом морали».

3.2. Интеллектуальные и когнитивные системы

Архитектура робота. «Обычная» интеллектуальная система управления выглядит так. На базовом, нижнем уровне реализуются элементарные двигательные функции и обработка сенсорной информации. Фактически это, как и у простейших животных, – рефлекторный уровень управления. На этом же уровне осуществляется усложнение реакций на внешние раздражители, формируются «библиотеки» развитых поведенческих программ. Но и такая система, пусть даже реализующая внешне нетривиальное поведение, попрежнему является условно-рефлекторной, хотя и способной даже к обучению. Далее формируется следующий уровень – над-стройка, ответственная за реализацию интеллектуальных функций. Речь идет о процедурах, способных производить логический вывод, осуществлять планирование и изощренный анализ внешних сигналов, распознавать сцены и принимать рациональные решения. Иногда вместо или совместно с этой интеллектуальной надстройкой создается иной, когнитивный уровень системы управления. На-

пример, в виде некоторой модели мира, в котором функционирует АИС. Эта модель представляет собой множество сущностей мира и семантических отношений между ними. Строго говоря, такая надстройка над базовым рефлекторным уровнем решает две основные задачи. Первая – это быть «отражением» тех сущностей, которые формируют базовый уровень (потребности, сенсорные сигналы, оценочные элементы и т. д.). Такое «отражение» позволяет реализовывать более тонкие и гибкие процессы управления, замещая или подменяя реальные сигналы базового уровня. Вторая задача – использовать существующие и формировать новые сущности модели мира. Этот уровень отвечает за рассуждения, получение новых знаний, выводы и прочие функции, которые принято относить к познавательным, интеллектуальным или разумным (если различать интеллект и разум и полагать, что разум – это способность устанавливать эмпирические закономерности [см., например: Крушинский 1986], а интеллект сводится к использованию познавательных процедур).

Эта когнитивная надстройка не является строго обязательной. Ее отсутствие (или отключение) не должно приводить к потере работоспособности АИС. В принципе, примерно так устроены и живые организмы. Здесь же важно, что когнитивный уровень может надстраиваться прочими уровнями, которые содержат не конкретные понятия и сущности, а более абстрактные, такие, например, как этические императивы. Иными словами, иерархия уровней управления реализует различные уровни адаптации АИС – от базовых безусловных реакций до элементов этического поведения. Ниже приведено условное изображение такой архитектуры. Здесь уровень R отвечает за условно-рефлекторную деятельность агента, а когнитивный уровень C может быть представлен в виде двух надстроек: уровень C1, описывающий механизмы взаимодействия агента с социумом (с себе подобными), и уровень C2, который отвечает за более абстрактные понятия, в том числе из области этики.

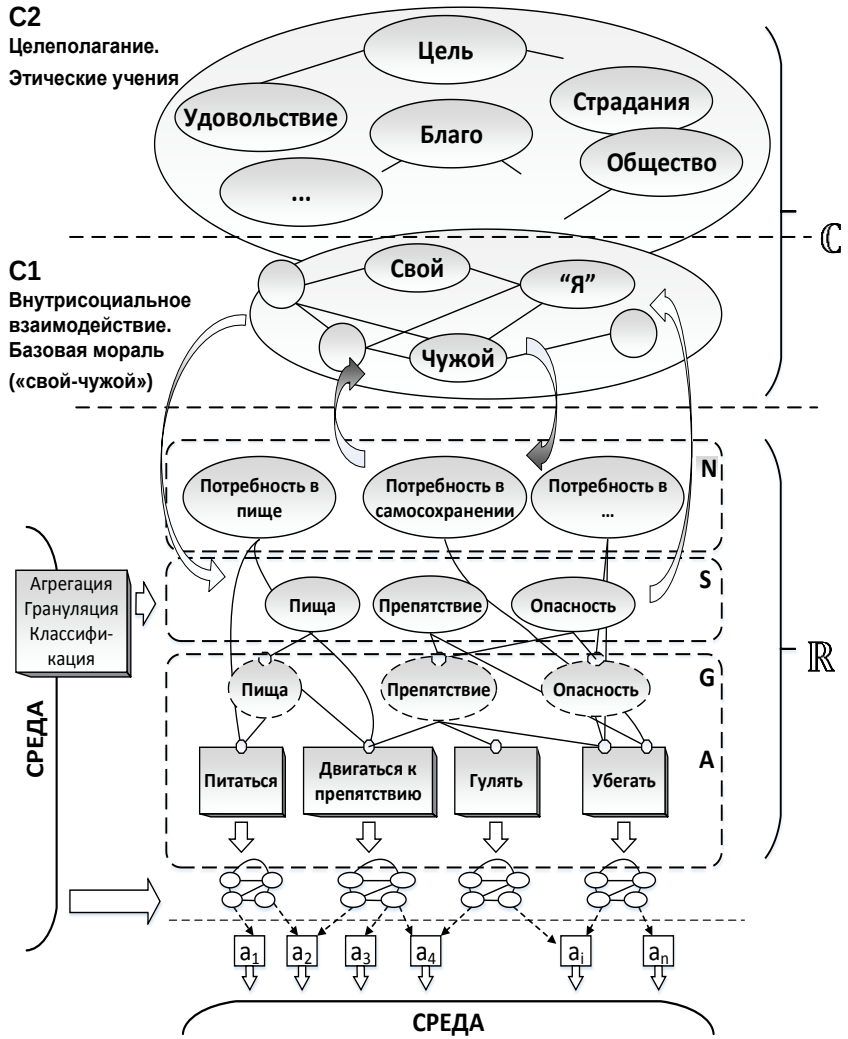

Рис. 2. Архитектура когнитивного робота

Этика как легковесная надстройка. Чем выше очередной уровень надстройки, тем она легковеснее и вариабельнее. Если грубое вмешательство в виде, например, изменения весов потребностей чревато потерей функциональности системы, то верхнеуровневые этические нормы крайне гибки и способны меняться многократно в зависимости от изменяющихся внешних условий. И это абсолютно естественно и закономерно. Представления о мо-

ральных категориях, о том, что такое хорошо/плохо и правильно/неправильно, могут меняться на протяжении жизненного цикла индивида (животного, робота) многократно, а вот удовлетворение базовых, физиологических потребностей – это вопрос выживания здесь и сейчас.

Очевидно, что при такой структуре, когда этические императивы образуют надстройку СУ, их «обход» – это сугубо технический вопрос.

3.3. Моральный робот-партнер

В качестве примера рассмотрим следующую задачу. Пусть у нас имеется коллаборативный робот, причем коллаборацию мы будем понимать как полноценное партнерство, а робота считать неким напарником человека, вместе с которым человек выполняет некоторую сложную и критически важную для человека задачу, например совместное патрулирование или охрану. Если среда, в которой действует пара человек – робот, является естественной, слабоформализуемой, недетерминированной, то основное требование к роботу-партнеру – это его автономность и, как следствие, интеллектуальность. Робот должен уметь анализировать сложную обстановку, планировать свое поведение, принимать решения. Фактически это означает, что его поведение неизбежно должно быть сродни поведению животного. Отсюда поведение робота может трактоваться с точки зрения удовлетворения его потребностей, основной из которых является потребность в самосохранении (а таковая должна быть, иначе робот быстро выйдет из строя). Итак, множество потребностей и поведенческих процедур, заложенных в робота и призванных удовлетворять его возникающие потребности, – это уровень простого животного, реализующего простое или сложное рефлекторное поведение. Это еще не полноценный партнер. Рассуждаем дальше.

Если роботу требуется реализовывать более сложное поведение, то надо уметь анализировать, предсказывать и оценивать результаты своих действий, общаться, в конце концов, с оператором. В сложной среде, решая сложные задачи, робот должен быть интеллектуальным. Но и такой интеллектуальный робот еще не партнер, так как вся его деятельность по-прежнему направлена на удовлетворение его потребностей. При этом роль оператора сводится к тому, чтобы манипулировать базовым поведением робота для выполнения прикладных задач (включаться в роль вида «свой-

чужой», формировать условия среды и т. д.). Партнером робот станет тогда, когда его поведение будет определяться еще чем-то, некими дополнительными оценочными механизмами, императивами или правилами. Эти правила и являются моральной составляющей поведения робота, образуя некую общую структуру – заложенное этическое учение. Суть любого этического учения заключается в определении приоритетов, оценок, целевых функций (что такое хорошо/плохо, правильно/неправильно и т. п.). Именно действия этих правил, наряду с базовыми рефлекторными, потребностными моделями поведения, и определяют поведение робота в целом. И тогда такой робот будет считаться полноценным партнером.

3.4. Коллизии. Главное – уметь объясниться

С другой стороны, неизбежно возникает коллизия между тем, что предписано правилами морали, и тем, что диктует рефлекторный уровень. В критической ситуации гипотетическое «этическое» азимовское правило вида «Необходимо подчиняться человеку» вступает в противоречие с правилом «Если опасность, то надо убегать» (правило подчинения действительно является этическим, так как определяет, наряду с прочими, целевую функцию существования робота). И мы вновь начинаем разбирать эти конфликты и противоречия, «подкручивать» коэффициенты важности правил и т. п., рискуя разбалансировать прежде целостное и устойчивое поведение робота, которое определялось системой понятных и логичных правил. На самом деле разрешение вопросов коллизий этического поведения робота-партнера осуществляется точно так, как это происходит в человеческом обществе. Вопрос не в том, какое действие или поведенческую процедуру реализует робот в той или иной ситуации. Вопрос в том, сможет ли он объяснить это действие с точки зрения заложенной в него этической схемы. Если да, то поведение робота в данной ситуации считается этичным, иначе – нет. Формально это означает: может ли робот построить цепочку рассуждений или доказательство того, что действие было обусловлено базовым этическим императивом. Ровно так поступают люди, находя оправдание своему поведению (ситуации, когда этические императивы превалируют над базовыми потребностями, в том числе – самосохранением, рассматривать не будем в силу их нетипичности).

4. Обсуждение и заключение

Резюмируем основные моменты.

Первое. Законы Азимова нельзя понимать буквально. Анализировать их, выискивать возможные примеры и контрпримеры, рассматривать возникающие коллизии – это, во-первых, не самое осмысленное занятие, а во-вторых, это вряд ли получится лучше и глубже, чем у самого Азимова: этому он и посвящал свои рассказы. Законы – это прежде всего повод задуматься о противоречивости и неработоспособности жесткой системы императивов, когда мы имеем дело с реальным, динамическим, недетерминированным и сложным (слабоформализуемым) миром; о роли эмоций; о том, насколько глубоко и подробно или, напротив, общезначимо следует ставить задачи для АИС.

Второе. Азимов ставит принципиально глубокий вопрос о моральной агентности робота. На самом деле, оставаясь в современной парадигме создания интеллектуальных или когнитивных архитектур СУ, мы можем говорить о том, что робот может быть полноценным моральным агентом, принимая решения, основываясь на нормах морали, неся ответственность и т. п. [см., например: Parthemore, Whitby 2013; Karpov 2020]. Важно, что это будет «обычный» моральный, агент, совсем не тот, о котором говорил Азимов. Обычный моральный агент принципиально ограничен и ситуативен при принятии решений. Абсолютная агентная моральность невозможна в силу того, что для этого необходимо просчитывать все возможные последствия действий и оценивать их с точки зрения морали, что накладно и нереализуемо технически. Если же у нас имеется некоторый реальный горизонт планирования оценок последствий, то вновь получится естественно ограниченный моральный агент. Кроме того, все равно не решаются проблемы коллизий при выборе морального решения. Единственный способ их разрешения – аргументировать или доказать моральность действия, а это уже совсем по-человечески (моральный компонент или императивы – легковесная вариабельная и потому малозначимая надстройка). Иными словами, моральный агент по образу и подобию человека может быть создан, но у Азимова как раз не об этом.

Третье. Нерешенной технической проблемой, которая требует серьезного исследования как с технической, так и с философской

стороны, является реализация этических императивов не в виде компонента когнитивной надстройки системы управления, а определение их на базовом уровне. Как создать жизнеспособного (адекватно функционирующего) робота, у которого моральные в человеческом понимании императивы выполняются рефлекторно, – это открытый вопрос. Если же создать робота, который руководствуется такими принципами на базовом уровне архитектуры, но при этом система останется функциональной, то мы столкнемся с еще более неприятной ситуацией, когда у агента будет совсем не человеческая мораль со всеми вытекающими гуманитарными и социальными последствиями.

Четвертое. Азимовские законы – это хороший повод к новым исследованиям. Важно и интересно, когда решение сугубо практических задач начинает требовать конструктивного переосмысления понятий, весьма далеких от технических и инженерных областей. Например, в том же рассказе «Улика» автор рассматривает вопрос разрешения коллизий законов и находит крайне неуклюжий, казалось бы, вариант их разрешения: «[Робот] сойдет с ума, если будет поставлен перед таким противоречием – нарушить букву Первого Закона, чтобы остаться верным его духу». На самом деле Азимов сумел здесь изящно обойти такой скользкий вопрос, как наличие у робота чувства вины. Обсуждение понятия вины с философской, психологической и прочих точек зрения выходит за рамки настоящей работы, однако с точки зрения технической наличие у «обычного», не азимовского робота такого чувства оказывается необходимым. Если согласиться с тем, что робот может быть моральным агентом, то неизбежным становится наличие механизма оценки моральности его – робота – поведения, создание таких механизмов, которые как минимум удобно называть совестью, чувством вины и т. п. При этом роль чувства вины сводится к реализации процедуры поощрения/наказания, определения штрафных воздействий и прочих оценочных механизмов, требуемых для процедуры обучения АИС. Вместе с тем роботы Азимова, будучи моральными агентами, обходятся без чувства вины, без обучения в смысле определения моральных последствий своих действий. «Обычный» робот действует по рефлекторно-ассоциативной схеме: в некоторой новой, неопределенной ситуации реализуется некоторая поведенческая процедура, далее оцениваются последствия. Если возникает

чувство вины (оценочный модуль выдает штраф), то в подобных ситуациях робот больше не будет выбирать эту процедуру, сформировалась соответствующая ассоциация. Азимовский же робот в случае противоречия просто перестает функционировать.

Это, разумеется, не единственная скрытая проблема азимовских законов, становящаяся актуальной сегодня с точки зрения необходимости решения вполне практических прикладных задач. Таковых проблем много, только надо постараться воспринимать образные рассуждения Азимова более глубоко.

Пятое. Превентивные ответы. Многолетнее обсуждение затрагиваемых в работе вопросов как в техническом, так и в гуманитарном сообществе привело к выделению следующих проблемных для восприятия мест (следует отметить, что в технической среде обсуждать эту проблематику легче – «технари» обычно убеждены, что им достаточно бытового понимания философских и психологических проблем). Приведем эти наиболее распространенные возражения.

(1) Критика «линейности» рассуждений. Это касается прежде всего реализации механизма эмоций робота, его шокирующей примитивизации. Еще раз отметим, что, во-первых, не надо путать эмоции и чувства; во-вторых, эмоции действительно откладываются на одной оси. А вот их проявления и анализ – это результат соотнесения общей оценки (собственно текущей эмоции) с конкретными поведенческими процедурами и актуальными потребностями. Когда же говорится о якобы «линейности», детерминированнности поведения (принятия решений) робота, то обычно игнорируют тот факт, что эмоционально-потребностная схема (кстати, она совсем не линейна) лишь запускает сложные поведенческие процедуры или их комплексы. Интеллектуальность, адаптивность, создание новых связей (прежде всего ассоциативных) и т. п., – все это находится внутри этих процедур, а также на верхнем, когнитивном уровне.

(2) Отсутствие определения основ морального поведения робота. Рассуждения о том, на каком учении должно базироваться поведение, несколько неконструктивны. С точки зрения ясности формулировок здесь хороша деонтология (робот должен поступать так-то, да еще исходя из неких приоритетов); с технической точки зрения, когда хочется уметь все взвесить и просчитать, удобна ак-

сиология (поступать, исходя из определения ценности человечества, человека, себя); гибкость и компромиссы обеспечивает утилитаризм и т. д. и т. п. Однако есть основания полагать, что все эти учения – лишь проекция поведения робота, взгляд с той или иной позиции в зависимости от ситуации. Главное, что все это никак не касается того, как устроена система управления робота, который выполняет свою основную задачу – функционирование в сложной естественной среде.

(3) Невозможность просчитать все последствия, определить опасности, угрозы и т. п. Причем с приведением массы убедительных примеров, хитроумных схем, содержащих логические ловушки и т. п. Мы уже ответили на этот вопрос, говоря об ограниченной и о ситуативной моральности. Ни на что большее физическая активная система претендовать не может. Решение, при котором робот становится созерцателем, исповедующим принцип «не навреди», вряд ли кому интересно.

Литература

Азимов А. Я, робот. М. : Эксмо, 2022.

Вааль Ф. де. Истоки морали: В поисках человеческого у приматов. М. : Альпина нон-фикшн, 2019.

Вейценбаум Д. Возможности вычислительных машин и человеческий разум. От суждений к вычислениям. М. : Радио и связь, 1982.

Гусейнов А. А., Апресян Р. Г. Этика. М. : Гардарики, 2000.

Дрейфус Х. Чего не могут вычислительные машины. Критика искусственного разума // What Computers Can’t Do: A Critique of Artificial Reason. М. : Прогресс, 1978.

Карпов В. Э. Эмоции и темперамент роботов. Поведенческие аспекты // Известия РАН. Теория и системы управления. 2014. № 5. С. 126–145.

Крушинский Л. В. Биологические основы рассудочной деятельности: Эволюционные и физиолого-генетические аспекты поведения. М. : Изд-во МГУ, 1986. Вып. 2.

Симонов П. В. Потребностно-информационная теория эмоций // Вопросы психологии. 1982. Т. 6. С. 44–56.

Симонов П. В. Мотивированный мозг. М. : Наука, 1987.

Asimov I. Robbie (Strange Playfellow) // Super Science Stories. 1940. September. Pp. 67–77.

Karpov V. Robot's Temperament // Biologically Inspired Cognitive Architectures. 2014. Vol. 7. Pp. 76–86.

Karpov V. E. Can a Robot Be a Moral Agent? // Artificial Intelligence. Lecture Notes in Artificial Intelligence (LNAI). 18th Russian Conference, RCAI 2020, Moscow, Russia, October 10–16, 2020. Proceedings / ed. by S. O. Kuznetsov, A. I. Panov, K. S. Yakovlev. Cham : Springer, 2020. Pp. 61–70.

Parthemore J., Whitby B. What Makes any Agent a Moral Agent? Reflections on Machine Consciousness and Moral Agency // International Journal of Machine Consciousness. 2013. Vol. 05. Pp. 105–129.

Plutchik R. The Nature of Emotions // American Scientist. 2001. Vol. 89. No. 4. Pp. 344–350.

Simonov V. P. Thwarted Action and Need – Informational Theories of Emotions // International Journal on Comparative Psychology. 1991. Vol. 5. No. 2. Pp. 103–107.