Karpov, V.E. (2020). Can a Robot Be a Moral Agent?. In: Kuznetsov, S.O., Panov, A.I., Yakovlev, K.S. (eds) Artificial Intelligence. RCAI 2020. Lecture Notes in Computer Science, vol 12412. Springer, Cham. https://doi.org/10.1007/978-3-030-59535-7_5

УДК 004.8:004.89:17

МОЖЕТ ЛИ РОБОТ БЫТЬ МОРАЛЬНЫМ АГЕНТОМ?[1]

В.Э. Карпов (karpov.ve@gmail.com)
Национальный исследовательский центр "Курчатовский институт", Москва,
МФТИ

Аннотация

Вопросы этически направленного проектирования интеллектуальных/автономных систем перешли сегодня в область нормативного регулирования. Если система должна принимать этически обусловленные решения, то она должна быть признана моральным агентом. В работе приводится перечень свойств, которыми должен обладать моральный агент и показывается, что искусственный агент не только может иметь такие свойства, но эти свойства обоснованы с технической точки зрения как адаптационные механизмы. В частности показано, что такие механизмы, как компонент “Я” в знаковой картине мира агента, наличие эмоционально-потребностной архитектуры, а также механизм сопоставления наблюдаемого контрагента с “Я” позволяют реализовать феномены социального обучения и такого свойства, как эмпатия.

Ключевые слова: моральный агент, эмоционально-потребностная архитектура, эмпатия, социальное обучение, подражательное поведение, этически обусловленное проектирование.

Введение

Этические проблемы искусственного интеллекта давно являются активно обсуждаемой темой. Более того, в последние годы эти вопросы перешли из разряда гуманитарных рассуждений в область технического регулирования. Речь идет, например, о глобальной инициативе IEEE по исследованиям в области этики ИИ. Результатом этих исследований должны стать технические регламенты, регулирующие разработку и внедрение систем ИИ с требованиями к их этическому поведению [IEEE, 2016]. Название документа очень примечательно: “Этически обусловленное проектирование” (“Ethically Aligned Design”). Другим показательным примером является отчет UNESCO по этике роботов, озаглавленный “Отчет COMEST по этике

[1] Работа выполнена при частичной финансовой поддержке РФФИ (проект № 17-29-07083-офи_м).

роботов” (COMEST – это Всемирная комиссия по этике научных знаний и технологий, World Commission on the Ethics of Scientific Knowledge and Technology) [UNESCO, 2017].

Большей частью рассуждения об этике интеллектуальных/автономных систем (И/АС) касаются различного рода угроз, социальных и экономических последствий их применения, этичности самих разработчиков и т.п. Нас же в этой работе интересует иной аспект этики И/АС. Этот аспект проявляется, когда мы рассматриваем И/АС как систему, автономно принимающую решения, критически важные для человека. В таком случае мы ожидаем, что эти решения будут соответствовать нашим представлениям об этичности. При этом способ применения этических механизмов при принятии решения не существенен. Например, этические соображения могут применяться для оценки того или иного возможного решения или действия. Оценка действия *Д*, совершаемого И/АС может определяться техническими, правовыми и моральными соображениями:

$$\begin{gathered} \textit{Оценка (Д)} = \\ \textit{Техническая\_оценка(Д)} + \textit{Правовая\_оценка(Д)} + \textbf{\textit{Моральная\_оценка}}\textit{(Д)} \end{gathered} \qquad (1)$$

В подобного рода рассуждениях здесь и далее вполне можно использовать в качестве иллюстрации различного рода вариации пресловутой проблемы вагонетки. С другой стороны, моральные соображения могут быть представлены как своего рода фильтр, задача которого – осуществить выбор среди множества альтернатив. Если принимаемое решение не может быть определено исходя из технических и правовых требований или ограничений, то должны быть применены некие дополнительные эвристики. Эти эвристики и есть правила этического характера. В этом заключается этичность поведения И/АС. Условно это можно изобразить так:

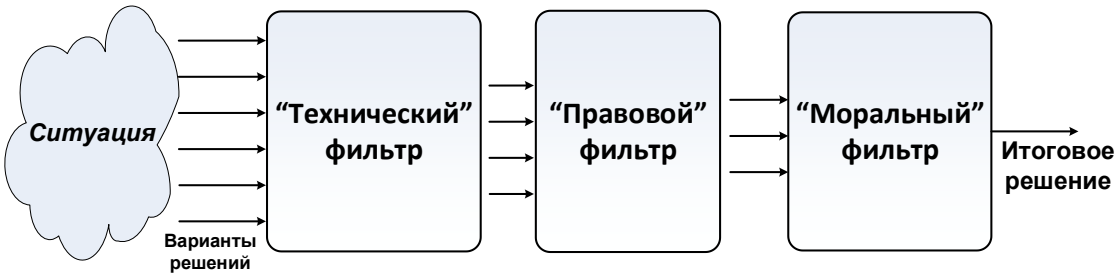

Рис. 1. Моральный выбор как способ разрешения неоднозначности

Предположим, что нам удалось формализовать положения моральной философии так, что они могут быть представлены некоторой системой правил (хотя это – сложная задача, требующая отдельного обсуждения). Но тогда неизбежно возникает следующая проблема. Если решение основано на моральных принципах (оно субъективно, противоречиво, слабо верифицируемо, нечетко и т.д.), то существует единственный способ повысить доверие к нему. Это – признание того, что решение было принято так называемым моральным агентом, т.е. некоторой сущностью, которой мы делегировали право применять соображения этического характера. Следовательно, ставится вопрос: существуют ли предпосылки для того, чтобы И/АС могла бы стать моральным агентом?

1. Моральный агент

Базовые определения сути морального агента обычно антропоцентричны и крайне лаконичны. Так, Parthemore и Whitby пишут: под “моральным агентом” мы понимаем любого агента, который надлежащим образом несет ответственность за действия [Parthemore, Whitby, 2013]. Далее часто добавляются рассуждения о том, что моральный агент действует в соответствии со своей ролью; говорится о свободе, как необходимом условии морального агента [Mayo, 1968]; понимании того, что такое хорошо и плохо и т.д. Более того, в [Parthemore, Whitby, 2013] говорится, что моральный агент обязательно является концептуальным агентом, то есть агентом, который обладает концепциями и использует их. В том числе – концепцией “Я”.

Мы не беремся обсуждать полный перечень свойств, которыми должен обладать моральный агент. Это – сугубо философская проблема. И эта проблема усугубляется отсутствием непротиворечивых конструктивных определений базовых понятий этики.

Нас интересует сугубо прикладной аспект создания И/АС. Поведение (принятие решений) И/АС должно отвечать нашим общим представлениям о поведении морального агента. При этом мы будем опираться на предположение, что моральным агентом может быть не только человек, а некая сущность вообще, в том числе – искусственный агент. Монополия человека в вопросах морали уже давно под сомнением (см., например, Ф. де Вааль, [Вааль де, 2019]), а мы делаем следующий шаг, отходя от биологического шовинизма вовсе.

Мы постулируем: множество проявлений свойств морального агента определяется тремя базовыми механизмами. Это (1) наличие у агента модели мира, в которой имеется компонент “Я” (познающий субъект), (2) механизм сопоставления наблюдаемого конспецифика или контрагента с “Я” и (3) наличие эмоционально-потребностной архитектуры системы управления нижнего уровня. При этом мы постараемся показать, что все эти компоненты имеют вполне практическое, реальное воплощение в технических системах. Рассмотрим далее, как эти механизмы позволяют реализовывать ряд феноменов поведения, присущих моральному агенту.

2. Феномены и механизмы

2.1. Эмоции и потребности

Начнем с нижнего уровня организации агентов. Роль эмоций в формировании этических норм и то, как эмоции определяют этику человеческого поведения, активно исследуются как философами, так и социологами [Neu, 2009], [Callahan, 1988], [Connelly, 1990]. Более того, Марвин Мински в своей работе [Minsky, 2006] предлагает рассматривать эмоции как другой способ мышления. Мы не станем обсуждать понятие эмоции в таких аспектах, не будем даже обсуждать вещи типа эмоционального интеллекта. Нас интересуют эмоции как сугубо физиологический механизм.

Есть все основания полагать, что эмоции (на физиологическом уровне) и темперамент (на психическом уровне) могут быть присущи технической системе как чисто прагматические механизмы, которые влияют на успех искусственного агента в сложных недетерминированных средах, см. [Karpov, 2014], [Карпов, 2014]. В этих работах эмоции рассматриваются как свойство системы управления, которая

способствует реализации таких функций, известных в психологии, как контрастное восприятие, стабилизация поведения, индикация состояния, работа в условиях неполноты информации и т.д. (см., например, [Ilyin, 2001], [Rai и др., 2018]).

Отметим здесь, что в архитектуре системы управления реакции системы, определяемые как эмоциональные, определяются контурами обратной положительной связи. Эти связи отвечают за оценку ситуации и определяют величину эмоционального состояния согласно Информационной теории эмоций П.Симонова [Симонов, 1982], [Simonov, 1991]:

$$E=f(N, p(I_{need}, I_{has})) \qquad (2)$$

где $\boldsymbol{E}$ – эмоция, ее величина и знак (качество); $\boldsymbol{N}$ – 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 агента средствах, (ресурсах), требуемых для удовлетворения актуальных потребностей. Здесь важно, что поведение агента (робота) определяется его потребностями и эмоциональным состоянием.

На Рис. 2 показан пример базовой эмоционально-потребностной архитектуры системы управления. «Эмоциональный» агент оснащен набором сенсоров и решает стандартную поведенческую задачу, используя несколько простых правил, таких как: "ЕСЛИ (голодный) ТОГДА (найти еду)", "ЕСЛИ (обнаружить препятствие) ТОГДА (убежать)" и т.п. Влияние эмоций на поведение агента реализуется как положительная обратная связь между выходными сигналами (текущими действиями или процедурами) и правилами поведения.

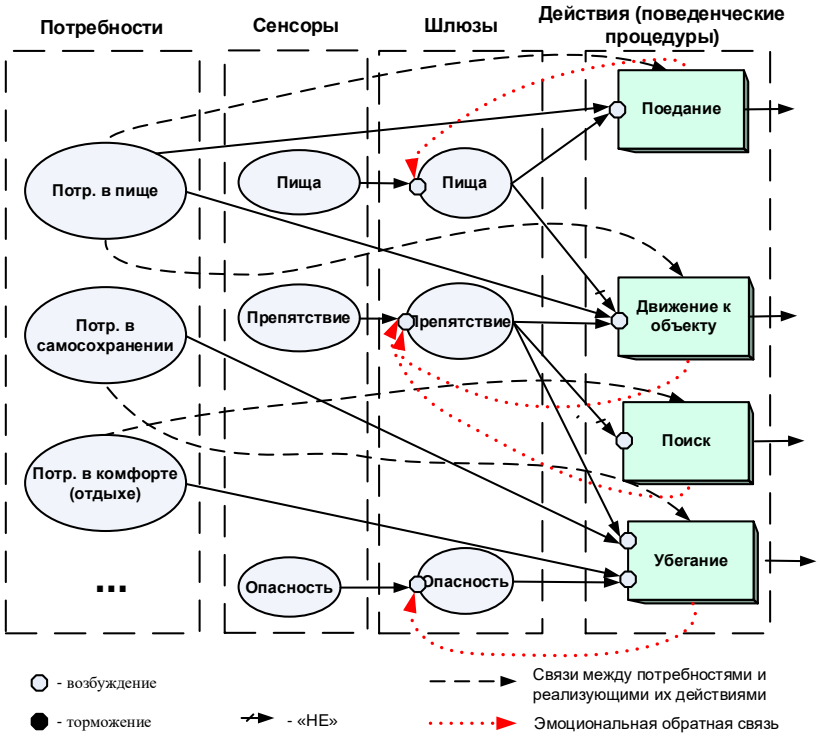

Рис. 2. Эмоционально-потребностная архитектура

Блок «Действия» – это набор поведенческих процедур. Каждая процедура активируется сигналами из блока «Потребности» и сигналами от специальных элементов «Шлюз». «Шлюз» – это элемент, который принимает прямые сигналы от датчиков и сигналы обратной связи от выходных элементов. Каждая выходная процедура имеет свой эмоциональный "вес". Этот сигнал является входным значением для элемента шлюза. Это означает, что положительные эмоции, связанные с некоторым действием («Поедание», «Поиск»,…), будут вызывать увеличение активности этого действия (проявление действия контура положительной обратной связи). Подчеркнем еще раз, что эмоционально-потребностная архитектура – это физиологический, базовый или рефлекторный уровень системы управления. Здесь основная функция эмоций – стабилизация поведения (принятия решений). Или, иной аспект, – определение мотивации поведения.

2.2. Модель мира, “Я”

Неотъемлемым атрибутом системы управления интеллектуального агента является наличие знаний об окружающем его мире. Если в эту модель мира добавляется компонент, называемый “Я” (субъект деятельности), то мы получаем то, что называется картиной мира (КМ) [Осипов и др., 2018]. В определенном смысле КМ может рассматриваться как некоторая надстройка над базовым стимул-реактивным уровнем. С архитектурной точки зрения, это тот компонент, который реализует воздействие на сенсорную систему, определяет значимость тех или иных потребностей и, таким образом, изменяет характер поведения системы, ее целеполагание и пр. Одной из наиболее эффективных моделей представления знаний в КМ являются знаковая или семиотическая модель. В этой модели основная сущность – знак – представлена четырьмя своими компонентами: имя n, перцепт p (образ, форма выражения), значение m (способ использования) и личностный смысл (цели, мотивы, личное значение) a. Однотипные компоненты образуют сеть, т.е. здесь мы имеем дело с четырьмя сетями.

Далее важны следующие допущения. Будем считать, что элементы системы управления снабжены еще одним, помимо возбуждающего или инициирующего, дополнительным подтверждающим входом. Это – вход для сигнала от вершины “Я”. Так, действие не будет активировано, если не будет подтверждающего сигнала от “Я”, интерпретируемого как "принадлежность" этого действия агенту. В определенном смысле это – чувство (ощущение) самости, т.е. отождествление или восприятие объекта как своего. Без такого ощущения в животном мире происходит рассогласование деятельности. Например, известно сложное психоневрологическое расстройство, называемое синдромом чужой руки. Одним из его клинических симптомов является наличие субъективных ощущений у пациента чужеродности конечности. С точки зрения семиотики это означает, что эти действия являются *значением* знака “Я”, т.е. вопрос обусловленности решается самым естественным образом.

Второе допущение заключается в том, что активизация какого-либо компонента знака влечет активизацию других его компонент. Кроме того, между одновременно активными вершинами сетей возникают ассоциативные связи.

2.3. Подражательное поведение

Такая модель достаточно естественным образом реализует такие феномены, как подражательное поведение и социальное обучение (обучение на основе наблюдения за другими). Приведем пример. Пусть агент знает, что объекты α_1 и α_2 съедобны, т.е. относятся к категории стимула *S* (еда) для действия *R* (съесть). Пусть далее агент наблюдает, что кто-то (конспецифик или контрагент) поедает объект *x*, который ранее не рассматривался агентом как съедобный. Тогда в результате этого наблюдения агент тоже отнесет этот объект к категории съедобных. Эта схема изображена на Рис. 3.

На Рис. За S_m – это компонент значения знака "съедобный объект", R_m и R_p – это компоненты значение и перцепт знака "поедание", $Self_m$ и $Self_p$ – это значение и перцепт знака “Я”.

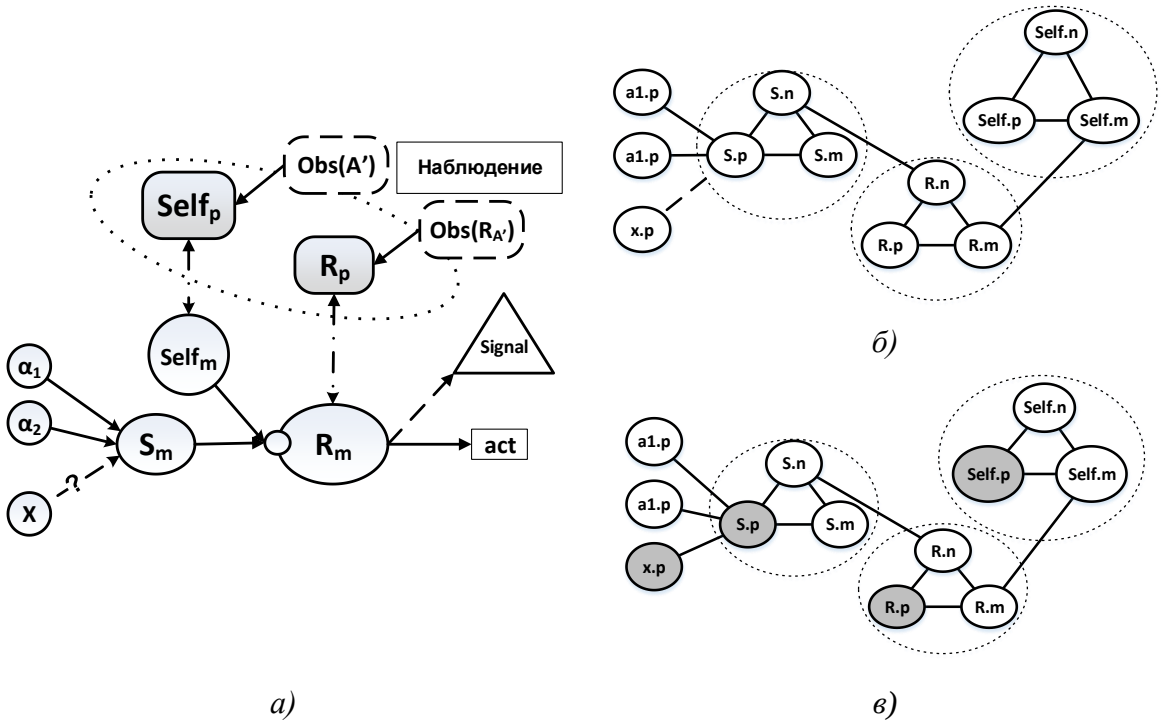

Рис. 3. Схема подражательного поведения: а) Концептуальная схема. Сплошные линии – установленные связи между элементами-значениями. Штрих-пунктирные линии – связи между компонентами знака: перцепт-значение, б) Начальное состояние ситемы: все вершины неактивны,имеются априорные связи между знаками и компонентами знака. Связь S.p-x.p является факультативной (выделена пунктиром). в) Ситуация наблюдения за действиями контрагента

Obs(A') и *Obs($R_{A'}$)* – это результат наблюдения: агент видит, что конспецифик *A'* выполняет действие *R*. Выполнение действия R_m активизирует некоторую моторную функцию (собственно поедание) *act*. При этом выполнение процедуры сопровождается выдачей вовне некоторого сигнала – *Signal*.

Итак, агент видит, что контрагент *A'* совершает некоторое действие *R* по отношению к объекту *X*. При этом *X* ранее не рассматривался субъектом, как определяющий фактор для стимула *S* (связь *X*-*S* не входила в личный опыт субъекта). Наблюдение за действиями контрагента приводит к активизации перцепта знака R_p.

Наличие связи перцепт-значение означает активизацию элемента-значения R_m: $Obs(R_{A'}) \rightarrow R_p \rightarrow R_m$. В то же время наблюдаемый контрагент сопоставляется с “Я”: активизируется перцепт $Self_p$, что приводит к активности $Self_m$.: $Obs(A') \rightarrow Self_p \rightarrow Self_m$.

Это – результат сопоставления контрагента с “Я”. Таким образом, в возбужденном состоянии оказываются все компоненты схемы: *R, S* и собственно объект наблюдения *X*. Между *X* и *S* формируется ассоциативная связь. Т.е. в ходе наблюдения объект *X* включается в поведенческий опыт анимата, причем делается это на основе наблюдения за поведением контрагента. Это и есть подражательное поведение. Добавление в эту схему оценочного элемента позволяет описать другой феномен – формирование рефлекторных реакций, также формируемых на основе наблюдения.

В этой схеме самым важным моментом является сопоставление субъекта с конспецификом (“Я” и другого), определение степени их близости. В природе это отождествление, возможно, сходно с тем, что называется родственным отбором, когда эволюционно выгодным становится поведение, определяемое степенью родства взаимодействующих особей, см., например, работу Уилсона [Уилсон, 2020].

2.4. Эмпатия

Под этим термином понимается способность к отзывчивости на эмоциональные состояния окружающих индивидов разной степени близости. Считается, что эмпатия определяет эмоциональную склонность к сотрудничеству и проявление альтруизма. Разумеется, предрасположенность индивида к эмпатии является необходимым, но не достаточным свойством для моральности индивида. Более того, эмпатия не является сугубо человеческим свойством. В этологии одним из обязательных механизмов, необходимых для формирования социального взаимодействия особей, является т.н. симпатическая индукция. Определение симпатической индукции тождественно определению эмпатии. Нас интересуют два аспекта эмпатии – механизм ее реализации, а также объект эмпатии.

Реализация механизма эмпатии возможна на том же принципе отождествления (или определения степени близости) наблюдаемого агента с “Я”. "Формула" эмпатии достаточно проста и определяется компонентами системы управления И/АС:

$$Эмпатия = \{Эмоции + Отождествление\ контрагента + Подражательное\ поведение\} \quad (3)$$

С объектом эмпатии дело обстоит несколько сложнее. В определенном смысле эмоции – это, прежде всего, способ интегральной оценки состояния индивида (см. [Симонов, 1982]). При этом мы полагаем, что имеется внешнее проявление эмоционального состояния агентов. Здесь же важно, что эмпатия – это основа для более высокого уровня управления, связанного с целеполаганием и планированием. В терминах моральной философии это означает действие золотого правила: либо реализовывать такой план действий, при котором контрагенту будет хорошо (увеличение уровня эмоционального состояния, позитивная формулировка правила: "поступай так, чтобы…"), либо сформировать план, не ведущий к появлению отрицательных эмоций контрагента (негативная формулировка: "не навреди"). В любом

случае эмоциональное состояние контрагента оказывает влияние на формирование мотива поведения, его цели. Это – основная роль эмпатии с технической точки зрения.

2.5. Характерные свойства морального агента

Далее мы тезисно приведем некоторые характерные свойства, обычно приписываемых моральным агентам. И посмотрим, насколько они важны и реализуемы технически.

Язык. Мораль не может существовать без символического языка. Однако мир животных прекрасно обходится без языка. То, что иногда называют языком, является сигнальной коммуникацией, т.е. внешним проявлением внутреннего состояния животного, см., например, работы Е.Н. Панова [Панов, 2014]. Роль сигнальной коммуникации, безусловно, велика. На этом построено множество механизмов социального поведения и взаимодействия. В том числе – и эмпатия, о которой мы говорили выше. Напомним, что выше мы усомнились в монополии человека на мораль.

Мотив. Вопросы мотивации поведения морального агента рассматриваются в моральной философии настолько разнообразно, что здесь можно найти любую удобную точку зрения (как, впрочем, и почти по всем иным вопросам). Например, очень удобным является такой взгляд Дж.Локка: личный интерес представляет собой единственный разумный мотив (цит. по [Darwall, 2007]). Если рассматривать мотив как опредмеченную потребность, то реализовать этого свойства можно механизмами прикладной семиотики.

Чувства. Если понимать под ними некий процесс, отражающий субъективное оценочное отношение к некоторым объектам, то те процессы, которые происходят в эмоционально-потребностной архитектуре, с полным основанием можно назвать чувствами (*feelings*). Робот действительно может испытывать чувства. Если же подразумевается, что моральный агент должен обладать так называемым "моральном чувством" (moral sentiments, [Smith, 2006]), то задача упрощается. Так, можно вовсе отказаться от рассмотрения этого свойства. При этом сославшись на Канта, который вывел из сферы морали чувства, считая их участие в мотивации деяний предпосылкой моральной ущербности последних [Апресян, Артемьева, Прокофьев, 2018]. Или счесть, согласно Ф. Хатчесону, что моральное чувство не может непосредственно мотивировать, а является откликом на мотив (цит. по [Darwall, 2007]). Там же говорится об утверждении Юма, что моральное чувство – результат действия более простых психологических начал – симпатии и ассоциации идей. Одобряются любые мотивы, ведущие к человеческому счастью. А хорошие последствия косвенно переживаются благодаря симпатии, ведущей к положительному чувству по поводу этого мотива.

Симпатия. Это явление также является следствием сопоставления наблюдаемого контрагента (даже не обязательно конспецифика) с “Я”. Естественно, сила симпатии зависит от степени близости контрагента. Следует отметить, что это может рассматриваться как прямое следствие организации знак-ориентированной картины мира. Возбуждение отдельных компонентов знака (перцепта, значения или личностного смысла) приводит к возбуждению его остальных компонентов, в том числе – и собственно имени "Я".

Ответственность. Моральный агент несет ответственность за свои действия. Но и здесь все может оказаться весьма простым. Согласно Parthemore & Whitby, моральный должен владеть определенными ключевыми понятиями и иметь способность в течение длительного периода взаимодействия между этим агентом и его социальной и физической средой соответствующим образом использовать эти понятия [Parthemore, Whitby, 2013]. Не помогает и такой взгляд на ответственность: "… Быть морально ответственным за что-то – скажем, за действие – значит заслуживать определенную реакцию – одобрение, порицание или что-то подобное – за его совершение" [Talbert, 2019].

Иногда к персональной ответственности за последствия решений добавляется требование самостоятельности решений, выражающихся в суждениях и действиях. При этом под самостоятельностью понимается то, что субъект не должен действовать в соответствии с заложенной кем-то программой и т.д. и т.п. Однако подобного рода рассуждения обычно быстро переходят в спекулятивную форму. Интересно, что такая постановка вопроса о границе между тем, что заложено природой, и свободой воли и самостоятельности, в технической интерпретации ставится крайне редко. Дело в том, что обычно есть четкое понимание того, что, с одной стороны, имеется некоторая фиксированная, априори заданная часть системы управления, а с другой – есть динамически изменяющиеся компоненты. Если вернуться к архитектуре анимата, то в нем четко выделен нижний – физиологический, фиксированный, рефлекторный уровень (на котором, кстати, работает эмоциональная часть системы управления). А надстройкой к нему является когнитивный уровень, представленный, например, семиотической системой.

Итак, мы можем констатировать, что, по меньшей мере, многие свойства морального субъекта могут быть присущи и искусственному агенту. Причем здесь речь не идет об имитации психических или когнитивных процессов, свойств сознания и т.п. Мы имеем дело с сугубо техническими решениями, которые призваны повышать адаптивные способности технического устройства. Другое дело, что эти решения – модели и механизмы – могут иметь интерпретацию в гуманитарных терминах.

Мы старательно обходили вопросы моральной философии. Рассуждения об утилитаризме, эволюционной этике и даже прагматизме не находятся в сфере наших компетенций. Мы пытались лишь поставить вопрос: если есть некий перечень свойств, которыми должен обладать моральный агент, то существуют ли причины, по которым мы не можем признать таковым искусственного агента – робота?

Заключение

Итак, все идет к тому, что мы делегируем интеллектуальной/автономной системе самостоятельно принимать решения, критически важные для человека. Если в ситуации выбора исчерпаны все соображения технического и правового характера, то остаются критерии морального характера. А доверие к такому "этическому" решению возможно лишь тогда, когда то, принимающее решение сущность является моральным агентом.

Подчеркнем еще раз, что все описанные выше механизмы вводились исключительно и только из соображений технической целесообразности, для решения

задачи создания эффективных адаптивных механизмов. Причем в три этапа, для решения трех классов задач. На первом этапе эти механизмы должны позволять техническому устройству целесообразно действовать в сложной, недетерминированной, динамической среде. На втором этапе решалась задача организации взаимодействия внутри группы агентов. Вплоть до появления форм социальной организации. При этом образование социумов агентов рассматривалось тоже как механизм адаптации. Третий этап – это задача целенаправленного управления поведения социумом. И вновь здесь были необходимы дополнительные механизмы адаптации, позволяющие социуму сохранять свою устойчивость. Одним из важнейших факторов стабилизации является наличие механизмов разрешения конфликтов внутри социума. А это и есть основная задача и суть морали.

Впрочем, моральной философии это тоже давно известно. Например, согласно Дробницкому, суть нормативной регуляции состоит в том, что "действие общественных закономерностей переходит в действия индивидуальных агентов", и таким образом "социальное целое воспроизводит себя через индивидуально-массовое поведение". Мораль представляет собой частный случай этого процесса, [Дробницкий, 2001], [Апресян, Артемьева, Прокофьев, 2018].

Сегодня идет интенсивное и достаточно успешное развитие когнитивных и социальных способностей интеллектуальных автономных систем. Однако в целом области этики поведения И/АС продвижение сопряжено с рядом сложностей, и основной проблемой является отсутствие конструктивных моделей, которых ждут исследователи от моральной философии. А их отсутствие приводит зачастую к тому, что разрабатываемые модели и методы остаются на уровне бытового, дилетантского понимания проблем морали.

Список литературы

[Апресян, Артемьева, Прокофьев, 2018] Апресян Р.Г., Артемьева О.В., Прокофьев А.В. Феномен моральной императивности. Критические очерки. М.: ИФ РАН, 2018. 196 с.

[Вааль де, 2019] Вааль де Ф. Истоки морали: В поисках человеческого у приматов. М.: Альпина нон-фикшн, 2019. Вып. 5. 376 с.

[Дробницкий, 2001] Дробницкий О.Г. Понятие морали: Историко-критический очерк // Дробницкий О.Г. Моральная философия: Избр. тр. / под ред. Р.Г. Апресян. М.: Гардарики, 2001. С. 11–344.

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

[Осипов и др., 2018] Осипов Г.С. и др. Знаковая картина мира субъекта поведения. М.: Физматлит, 2018. 264 с.

[Панов, 2014] Панов Е.Н. Эволюция диалога. Коммуникации в развитии: от микроаргонизмов до человека. М.: Языки словянской культуры, 2014. 400 с.

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

[Уилсон, 2020] Уилсон Э. Эусоциальность: Люди, муравьи, голые землекопы и другие общественные животные. М.: Альпина нонфикшн, 2020. 158 с.

[Callahan, 1988] Callahan S. The Role of Emotion in Ethical Decisionmaking // Hastings Cent. Rep. 1988. Т. 18. № 3. С. 9.

[Connelly, 1990] Connelly J.E. Emotions and the process of ethical decision-making. // J. S. C. Med. Assoc. 1990. Т. 86. № 12. С. 621–3.

[Darwall, 2007] Darwall S. The Foundations of Morality: Virtue, Law, and Obligation // The Cambridge Companion to Early Modern Philosophy / под ред. D. Rutherford. Cambridge: Cambridge University Press, 2007. С. 221–249.

[Ilyin, 2001] Ilyin E.P. Emotions and Feelings (in Russian). Saint-Petersburg: Piter, 2001. 752 с.

[Karpov, 2014] Karpov V. Robot's temperament // Biol. Inspired Cogn. Archit. 2014. Т. 7. С. 76–86.

[Mayo, 1968] Mayo B. The Moral Agent // Royal Institute of Philosophy Lectures. , 1968. С. 47–63.

[Minsky, 2006] Minsky M. The emotion machine : commonsense thinking, artificial intelligence, and the future of the human mind. Simon & Schuster, 2006. 387 с.

[Neu, 2009] Neu J. An Ethics of Emotion? : Oxford University Press, 2009.

[Parthemore, Whitby, 2013] Parthemore J., Whitby B. What makes any agent a moral agent? Reflections on machine consciousness and moral agency // Int. J. Mach. Conscious. 2013. Т. 05.

[Rai et al., 2018] Rai M. и др. Extraction of Facial Features for Detection of Human Emotions under Noisy Condition // 2018. № September. С. 49–62.

[Simonov, 1991] Simonov V.P. Thwarted action and need – informational theories of emotions // Int. J. Comp. Psychol. 1991. Т. 5. № 2. С. 103–107.

[Smith, 2006] Smith A. The Theory of Moral Santiments. : MetaLibri, 2006. Вып. 6. 322 с.

[UNESCO, 2017] UNESCO. Report of COMEST on Robotics Ethics // 2017. С. 64.

CAN A ROBOT BE A MORAL AGENT?

Valery E. Karpov (karpov.ve@gmail.com)
National Research Centre "Kurchatov Institute", Moscow,
Moscow Institute of Physics and Technology (State University), MIPT

Issues of the ethically aligned design of intelligent/autonomous systems have now moved into the fields of normative and technical regulation. If a system must make ethically determined decisions, then it must be recognized as a moral agent. This paper provides a list of the properties of a moral agent and shows not only that an artificial agent can have such properties, but also that they are technically determined as manifestations of adaptive mechanisms. In particular, it is shown that mechanisms such as the presence of the "I"

component in the sign-oriented picture of the agent's world, the presence of an emotional-needs architecture, and the mechanism for comparing the observed conspecific with the "I" make it possible to realize the phenomena of social learning and a property such as empathy.

Keywords: moral agent, emotional-needs architecture, empathy, social learning, imitative behavior, ethically aligned design.