УДК 004.8: 17
DOI:10.38028/ESI.2022.28.4.018

К вопросу о формализации этики поведения коллаборативного робота

Карпов Валерий Эдуардович[1], Королева Мария Николаевна[2]
[1]НИЦ «Курчатовский институт»,
[2]МГТУ им. Н.Э. Баумана,
Россия, Москва, *maria.svyatkina@gmail.com*

Аннотация. В работе рассматриваются вопросы создания механизмов, позволяющих коллаборативному роботу считаться моральным агентом. Показано, что в условиях принципиально неустранимых противоречий и неполноты использование логических схем и различных неклассических логик не решают проблему формализации моральных учений и этики поведения. На примере показано, что задача моральных схем или систем суждений заключается не только и не столько в том, чтобы определить или оценить действие (поведение) агента, а в том, чтобы реализовывать объяснительный компонент поведения. В качестве этических теорий, закладываемых в основу поведения робота, рассматриваются утилитаризм в понимании И. Бентама и Дж. С. Милля в комбинации с аксиологической системой правил гедонизма. Для качественного анализа применяются когнитивные карты. В работе сформулированы три фундаментальные утверждения морального поведения робота: 1) поведение (выбор действия, реализация поведенческой процедуры) агента определяется текущими актуальными потребностями и состоянием его системы восприятия; 2) в основе моральности поведения лежит результат сопоставления конспецификов с «Я», т.е. определение степени «свой-чужой»; 3) схема моральных представлений нужна для того, чтобы определить значимость потребностей и особенности системы восприятия, а главное – обосновать (оправдать) выбранное агентом поведение.

Ключевые слова: этика, коллаборативный робот, неклассические логики, моральный агент, анимат, эмоционально-потребностная архитектура, эмпатия

Цитирование: Карпов В.Э. К вопросу о формализации этики поведения коллаборативного робота / В.Э. Карпов, М.Н. Королева // Информационные и математические технологии в науке и управлении. – 2022. – № 4(28). – С. 223-233. – DOI:10.38028/ESI.2022.28.4.018.

Введение. Одним из векторов развития популярного междисциплинарного направления «Объяснимый искусственный интеллект» является формирование системы методов и технологий для поддержки пользователей (естественных агентов) в плане понимания ими своих искусственных интеллектуальных партнёров и формирования у людей доверия к их решениям. Примером партнёрских систем является коллаборативная робототехника, в которой рассматриваются вопросы взаимодействия, понимания, объяснения и интерпретации поведения системы «человек-робот». Эти проблемы подробно рассмотрены в работе В.Б. Тарасова [1]. При этом взаимное понимание и доверие во многом определяются этическими аспектами взаимодействия. По сути, этическая оценка поведения говорит нам о субъекте – насколько он определяется как свой или чужой, можем ли мы ему доверять, действуем ли мы с ним в рамках некой единой системы ценностей. В данной работе, продолжая исследования В.Б. Тарасова, остановимся на вопросе формализации этики поведения искусственных агентов в партнёрских системах. Коллаборативный робот – это робот, непосредственно взаимодействующий с человеком при выполнении совместных работ. Отсюда – требование к тому, чтобы поведение робота отвечало этическим представлениям человека-оператора. Причина в том, что функционирование осуществляется в естественной, недетерминированной среде, когда технические регламенты и ограничения не всегда позволяют роботу выбирать требуемое поведение. Существуют ситуации, когда рациональный выбор действия не будет предпочтительным с точки зрения человека-оператора. Второй аспект необходимой моральности поведения заключается в том, что при тесном контакте человек наделяет робота-партнёра свойствами морального агента, что требует необходимости принятия решений роботом, исходя из неких представлений о морали.

В работе мы, согласно Р.Г. Апресяну [2], будем полагать термины «этика», «мораль» и «нравственность» синонимами. Кроме того, согласно тому же Р.Г. Апресяну, мы будем считать, что основная задача морали – это разрешение конфликтов. Основным же конфликтом будем полагать наличие неоднозначности при принятии решений (конфликтное множество правил в инженерии знаний). Здесь же отметим, что для решения нашей задачи – формализации этики поведения – не достаточен подход В. Лефевра из его «Алгебры совести» [3]. Этическая система строится им на наборе правил перехода от элементарных ценностей к их комплексам, которые, в свою очередь, определяются только бинарными оценками «добро-зло». В конечном счете, такая этическая система дает лишь бинарную оценку действия, но не способна объяснить и оправдать поведение.

На основе очевидных для гуманитарных сфер проблем, перечислим особенности формализации этики:

1. Слабо и плохо формализуемые понятия.
2. Неизбежные противоречия.
3. Изолированные участки формализации: схема понятий, причинно-следственные связи не всегда образуют целостную систему.
4. Многое «объясняется» примерами, частными случаями, прецедентами (язык притч).

В этой работе мы рассмотрим один из подходов к формализации этического аспекта поведения робота с сугубо прагматической, технической точки зрения.

1. Постановка инженерной задачи. Пусть перед неким гипотетическим инженером стоит задача создания искусственного агента (для наглядности – робота), который должен не только функционировать в естественной человеческой среде, но и выполнять вместе с человеком некоторую работу, т.е. функционировать в социуме самым непосредственным образом. Предположим, что наш инженер смог определить базовую функциональность робота, определив множество его поведенческих программ, например, по выполнению тех или иных технологических операций. Далее, наш инженер понимает, что сложная среда функционирования подразумевает неоднозначность реакций робота. Неоднозначность обычно разрешается путем введения оценок возможных действий, а вот сами оценки определяются техническими, правовыми и моральными соображениями. Если принимаемое решение не может быть однозначно определено, исходя из технических (нажать на тормоз) и правовых (причинение минимального вреда) требований, то должны быть применены некие дополнительные фильтры в виде эвристик (сделать так, чтоб было «хорошо»). Такими эвристиками и являются моральные соображения. Условно это можно изобразить так:

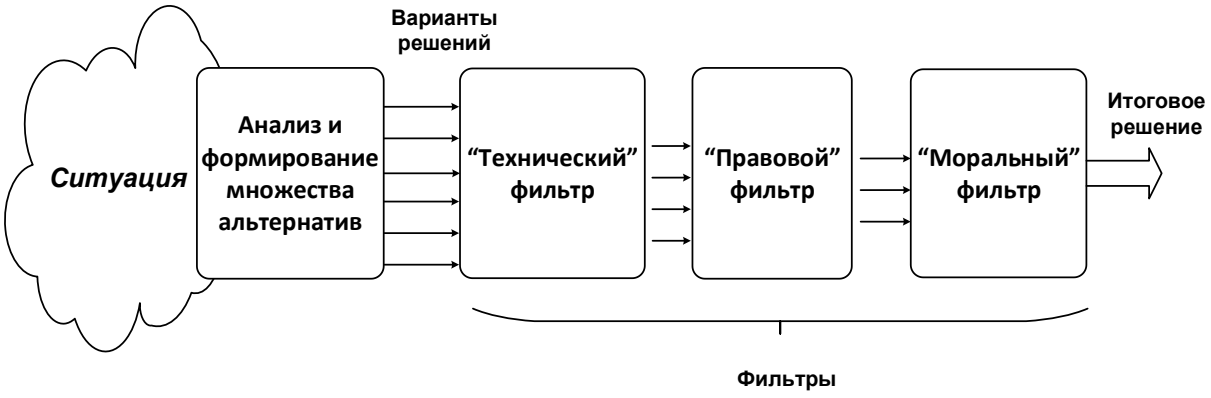

Рис. 1. Моральный выбор как способ разрешения неоднозначности (по мотивам [4])

Итак, итоговая оценка некоторого действия *Д* может быть представлена в виде:

$$Оценка\ (Д) = Техническая\_оценка(Д) + Правовая\_оценка(Д) + \mathbf{Моральная\_оценка}(Д) \quad (1)$$

Итак, предположим, что среди множества этических теорий, закладываемых в основу поведения робота, был выбран утилитаризм в понимании И. Бентама и Дж. С. Милля [5]. Это понимание несколько ближе к инженерному подходу в силу того, что, во-первых, речь у них идет о рационализации хотя бы экономического, а не сугубо гуманитарного поведения, а, во-вторых, Дж. С. Милль, прежде всего, – логик. К тому же, именно поведение – основной аспект утилитаризма. Итак, пусть имеется некое множество постулатов утилитаризма:

У1. Счастье, как получение удовольствия и избавление от страданий, является смыслом деятельности человека (под счастьем подразумевается удовольствие и отсутствие боли).

У2. Полезность человека и его дел для общества – это самый значимый критерий оценки всех явлений.

У3. Главный критерий нравственности – принцип полезности в виде стремления к достижению счастья для наибольшего числа людей.

У4. Удовольствие может проистекать из дружбы и симпатии.

У5. Основная цель развития человечества – стремление к расширению всеобщей пользы через гармоничное сочетание интересов (счастья) каждого индивидуума с общими интересами.

Полнота этого перечня, корректность формулировок для нас принципиально не существенна. Очевидно, что для инженера здесь явно не хватает оценочных суждений. Нужны явные критерии, системы оценок или хотя бы то-то такое, что оперировало бы некими ценностями. К счастью, наш гипотетический инженер слышал что-то о существовании теории ценностей – аксиологии. Добавим к постулатам У1-У4 что-нибудь из таких аксиологических учений, например, гедонизм. Тем более что гедонизм – это основа утилитаризма. И тогда мы получим такую дополнительную систему правил [6]:

Г1. Смысл человеческих действий и основа счастья – стремление к удовольствию и отвращение от страдания (эквивалентно У1).

Г2. Следует поступать всегда так, чтобы можно было непосредственно удовлетворять свои потребности и испытывать как можно большее наслаждение.

Г3. Запреты и предписания (влияние социума) – это препятствие к достижению удовольствия.

Наши рассуждения о морали будут иметь качественный характер. Изобразим эти правила-постулаты в некотором наглядном, качественном виде, пригодном, к тому же, для последующего анализа.

Когнитивные карты. Одним из механизмов качественного анализа являются когнитивные карты. Далее мы будем опираться на работы О.П. Кузнецова и А.А. Кулинича, см., например, [7, 8]. Когнитивная карта – это ориентированный граф, ребрам которого поставлены в соответствие веса. Вершины графа соответствуют факторам (концептам), определяющим ситуацию, ориентированные ребра – причинно-следственным связям между факторами. Веса определяют силу влияния этих факторов. Положительный вес означает, что увеличение фактора-причины приводит к увеличению значения фактора-следствия, отрицательный – к соответствующему уменьшению.

Если веса графа принимают значение +1 и −1, то мы имеем дело со знаковым графом. Важная задача в анализе таких графов – это анализ его циклов. Положительный цикл – это контур положительной обратной связи. Увеличение значения некоторого фактора в этом цикле приведет к его дальнейшему неограниченному росту, т.е. к потере устойчивости. Отрицательный цикл противодействует отклонениям от начального состояния и способствует его устойчивости. Знак цикла определяется знаком произведения его ребер.

Когнитивная карта, соответствующая правилам У1-У5 и Г1-Г3, изображена на рис. 2.

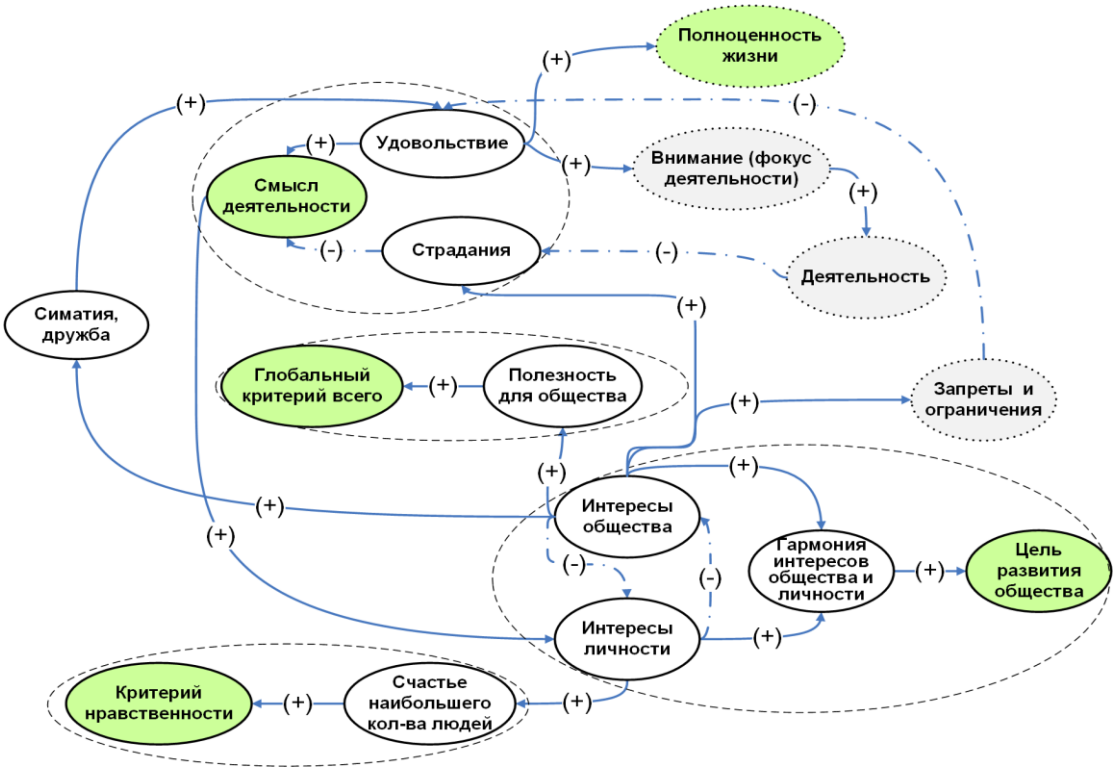

Рис. 2. Фрагмент когнитивной карты этической системы

Пока не будем обращать внимание на то, что на этой карте есть «висящие» вершины-понятия, а ряд вершин вполне можно объединить или, напротив, следует представить в более подробном виде. Наша задача – определить суть технологии.

На этой карте есть некие «противоречия». Например, говорится, что влияние общества (на схеме «Интересы общества») в виде запретов и ограничений отрицательно, т.е. уменьшает удовольствие (Г3). С другой стороны, влияние общества в виде симпатии/дружбы повышает удовольствие.

Формально мы получаем утверждение вида:

$$\textit{Интересы общества} \vdash \textit{Удовольствие}$$
$$\textit{Интересы общества} \vdash \neg\textit{Удовольствие}$$

Или, если полагать, что здесь вполне можно воспользоваться правдоподобностью, интерпретировать это как $\{\textit{Интересы общества}, \neg\textit{Интересы общества}\} \vdash \textit{Удовольствие}$. Здесь $\vdash$ – знак логической выводимости.

Разумеется, здесь есть и более серьезные сложности. Например, определение семантики используемых сущностей, определение того, как это будет представлено в картине мира робота и т.п., но пока ограничимся лишь проблемой явных внешних противоречий.

Очевидно, что появление высказывания $\{\alpha, \neg\alpha\} \vdash \beta$ приводит нас к выводу о необходимости применения специального класса логических исчислений, в которых логический принцип *«из противоречия следует все что угодно»* (*"ex contradictione (sequitur) quodlibet"* – *ECQ*), не имеет места. Это – область т.н. *паранепротиворечивых логик*. Мы не беремся утверждать, что при описании философских конструкций достаточно лишь паранепротиворечивых логик. Возможно, здесь могут потребоваться *подлинно паранепротиворечивые логики*, в которых отвергается не только принцип *ECQ*, но отвергается и закон непротиворечия (*NC*): $\neg(\alpha \wedge \neg\alpha)$, см. [9].

Существуют различные способы опровержения и ограничения принципа *ECQ*, более того, паранепротиворечивых логик бесконечно много, однако среди их множества нас интересуют т.н. *релевантные логики*. В релевантной логике доказуема не каждая формула, глав-

ный знак которой – импликация, а антецедент противоречив. Существуют различные семантические подходы, показывающие паранепротиворечивость релевантных логик, например, *семантика возможных миров* С. Крипке. В классическом понимании возможных миров полагается [10]: 1) ни в каком мире w, никакая пропозициональная переменная α не встречается одновременно со своим отрицанием (условие непротиворечивости); 2) во всяком мире w, любая пропозициональная переменная α встречается либо сама, либо с отрицанием (условие полноты). При этом доказуемыми являются те и только те формулы, которые являются истинными во всех возможных мирах. Здесь главным вопросом является интерпретация отрицания. С каждым миром w ассоциируется мир w^*, при этом α истинно в w тогда и только тогда, когда α ложно не в w, а в w^*, т.е. если α истинно в w, но ложно в w^*, то $(\alpha \wedge \neg\alpha)$ истинно в w [11].

На самом деле, использование какого-либо вида паранепротиворечивой логики не сможет решить нашу проблему. Дело в том, что нас не интересует, каким образом можно переформулировать нашу систему этических представлений, выраженных в виде ряда правил-высказываний. Она противоречива принципиально. Модификация моральной схемы, введение новых сущностей, разрешение семантических неоднозначностей – это уход от проблемы, точнее, это – просто перенос проблемных вопросов на другой уровень. Не решит проблему формализации и переход к вероятностным или нечетким рассуждениям. Нам необходимо оставаться в рамках системы явных правил-высказываний. Это связано с тем, что:

1. Схема принятия этического решения или этической оценки ситуации должна быть предельно наглядна, интерпретируема и, главное, объяснима.
2. Все модификации, расширения моральной схемы осуществляются именно в форме явных языковых высказываний.
3. Схема должна быть компактной в силу необходимости ее анализа на каждом шаге принятия решения или выбора действия агентом.

Рассмотрим некую модификацию (или развитие) схемы, приведенной на рисунке 2. Будем полагать, что наша задача – определить, какое действие (или поведение) должен реализовать агент. Предположим, что среди множества возможных действий в некоторый момент времени агенту доступны два – *«Отдыхать»* и *«Работать»* (рис. 3).

Действие *«Работать»*, согласно фрагменту схемы, приведет, с одной стороны, к получению удовольствия, а с другой – к его отрицанию через ограничения, накладываемые социумом. Формально это означает наличие нескольких путей из вершины *«Работать»* в вершину *«Удовольствие»* с разными знаками. Или:

$$Работать \vdash Удовольствие$$
$$Работать \vdash \neg Удовольствие$$

Мы не будем оценивать эти пути, вводя различные веса переходов, коэффициенты значимости и т.п. Повторим, что это в конечном итоге – тоже уход от проблемы. Нам надо понять, как действовать именно в условиях неустранимых противоречий. В данном случае, агенту необходимо определить, что делать – отдыхать или работать. И здесь мы приходим к выводу, что на самом деле задача наших моральных схем, систем суждений и т.д. совершенно не в том, чтобы определить или оценить действие (поведение) агента. Как будет показано ниже, задача моральной системы лишь в том, чтобы реализовывать объяснительный компонент поведения.

Рассмотрим далее вопрос о том, что лежит в основе поведения искусственного агента. Будем далее иногда вместо совершенно общего и абстрактного термина агент использовать термин *анимат*, подчеркивая, что нас интересует некая модель, претендующая на описание поведения если не человека, то животного [12].

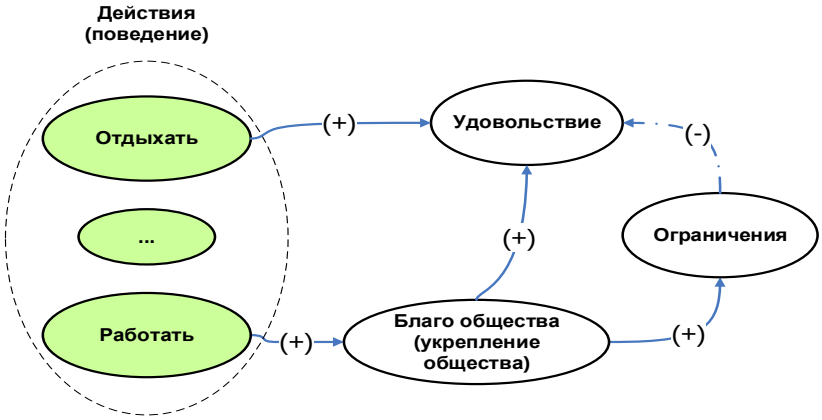

Рис. 3. Выбор действия в условиях противоречий

2. Базовое поведение анимата. Поскольку мы говорим о поведении искусственного агента – робота, то неизбежно встает вопрос об архитектуре его системы управления, причем, начиная с самого нижнего уровня. Будем полагать, что среди множества возможных архитектур мы остановились на следующей (рис. 4).

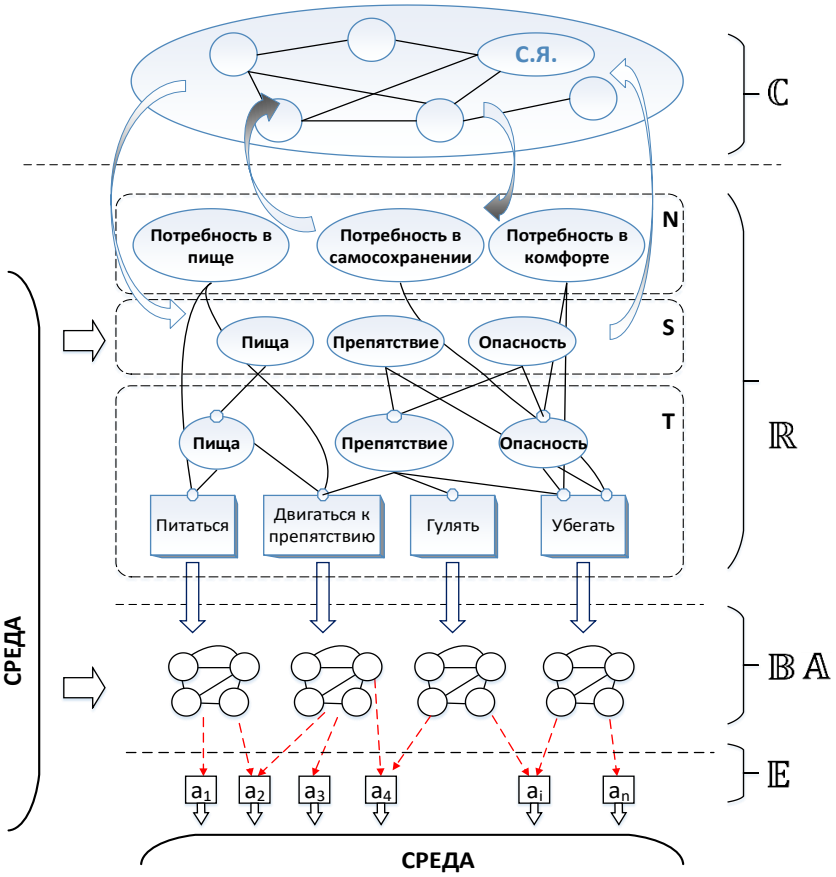

Рис. 4. Схема обобщенной архитектуры системы управления анимата

Непосредственное общение со средой (датчики, сенсорные подсистемы, эффекторы) осуществляется на уровне $\mathbb{E}$. Уровень $\mathbb{BA}$ – это уровень поведения (от простого до сложного, вплоть до поведенческих комплексов). Поведенческий уровень может быть реализован на базе, например, мета-автоматной логики. $\mathbb{R}$ – это основной регуляторный уровень. На этом уровне осуществляется запуск поведенческих процедур, причем выбор того или иного поведения основан на имеющихся у агента актуальных потребностей (N), информации от сенсорной системы (S) и эмоционального состояния агента. Механизм эмоций отвечает за устойчивость поведения агента, контрастирование его восприятия и за кратковременную память. Это то, что лежит в основе стабилизирующей части (T). Фактически, на уровне $\mathbb{R}$ мы имеем дело с т.н. эмоционально-потребностной схемой.

Уровни $\mathbb{E}, \mathbb{BA}$ и $\mathbb{R}$ относятся к рефлекторному – нижнему – уровню управления. На рисунке 4 для наглядности условно приведены обозначения потребностей и абстрактных сигналов сенсорной системы (подсистемы восприятия).

На самом верху располагается когнитивный уровень $\mathbb{C}$, часть, ответственная за отображение сенсорных и потребностных элементов. Этот уровень называется моделью мира, здесь формируется мотивация и целеполагание, осуществляется планирование и т.п. Реализуется этот уровень, например, семиотическими моделями. Формально работа такой системы управления агента может быть определена следующим образом.

Пусть Y_{eff} – вектор выходных эффекторных воздействий анимата на среду. Его формирование определяется результирующим выходом последовательности $n+1$ мета-автоматов M^k, $k=0...n$.

$$Y_B \overset{a}{\Rightarrow} M_i^n \overset{a}{\Rightarrow} M_j^{n-1} \overset{a}{\Rightarrow} \dots \overset{a}{\Rightarrow} M_k^0 \overset{a}{\Rightarrow} Y_{eff}.$$

Здесь Y_B – вектор активации мета-автомата, определяющего выбранное актуальное поведение, $\overset{a}{\Rightarrow}$ – отношение непосредственного автоматного управления, задаваемое соответствующими функциями выхода автоматов λ: $Y^k = \lambda(Q^k, S^k),\ S^k = X \cup f(Q^{k-1})$. Здесь Q^k, S^k – состояния и входной алфавит мета-автомата уровня k соответственно. При этом S^k определяется как объединение внешних сигналов (сигналов среды) и текущего состояния подчиненного автомата уровня $k-1$.

Вектор активации поведения Y_B задается функционированием регуляторного уровня $\mathbb{R}$:

$$Y_B = \mathbb{R}(S, N, G),$$

где S, N, G – сенсорные сигналы, потребности и вентильные элементы эмоционально-потребностной схемы соответственно. Подробное описание «эмоционального» компонента систему управления выходит за рамки настоящей работы. Отметим здесь лишь то, что в его основе – информационная теория эмоций П. Симонова [13, 14]. Согласно ей, интегральная оценка ситуации определяется оценкой баланса между необходимыми и имеющимися средствами удовлетворения актуальных потребностей. Оценочная, качественная формула Симонова выглядит так:

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

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 агента средствах, (ресурсах), требуемых для удовлетворения актуальных потребностей. «Технически» оценка (2) формируется контурами обратной связи между действием, направленным на удовлетворение актуальной потребности и интегрирующим сенсорные и потребностные сигналы элементы-вентили G.

Задача когнитивного уровня $\mathbb{C}$ – формировать целенаправленное поведение агента, при этом воздействие на регуляторный уровень осуществляется опосредованно, через элементы G (сенсорика и потребности – это фиксированный, «аппаратный» уровень архитектуры): $G = \mathbb{C}(S, N)$. В итоге получаем:

$$\mathbb{R}\big(S, N, \mathbb{C}(S, N)\big) \Rightarrow^M Y_{eff}. \qquad (3)$$

Здесь $\Rightarrow^M$ – транзитивное замыкание отношения автоматного управления $\overset{a}{\Rightarrow}$.

Все эти рассуждения необходимы лишь для того, чтобы обосновать наличие конструктивной схемы управления, определяющей поведение анимата. Итак, мы можем сформулировать достаточно очевидное утверждение:

Утверждение №1. Поведение (выбор действия, реализация поведенческой процедуры) анимата определяется текущими актуальными потребностями и состоянием его системы восприятия.

Перейдем далее к основаниям морального поведения.

3. Основания морального поведения. Наиболее явным механизмом, имеющим теснейшую связь с моральными аспектами поведения, является эмпатия – способность к отзывчивости на эмоциональные состояния окружающих индивидов разной степени близости (в животном мире это свойство называется симпатической индукцией). Эмпатия определяет склонность к сотрудничеству и проявлению альтруизма. Реализуется этот механизм на основе процедуры отождествления (определения степени близости) наблюдаемого другого агента с образом «Я». При этом удобной формой представления и описания процедуры сопоставления является использование семиотической модели, в которой понятие «Я» рассматривается как знак, обладающий смыслом, значением и образом (перцепт знака). Будем называть далее наблюдаемых других агентов конспецификами, т.е. представителями одного вида. При этом предполагается, что аниматы способны наблюдать и идентифицировать эмоциональное состояние конспецификов. Очевидно, что эта семиотическая модель реализуется на уровне $\mathbb{C}$ рассмотренной выше архитектуры анимата.

Определяемая степень близости с конспецификами – это основа для разделения «своих» и «чужих». Здесь нам приходится вернуться к т.н. золотому правилу морали [15], которое задает целевую функцию морального поведения. Это правило может быть дано в позитивной (поступать по отношению к другим так, как желаете, чтобы поступали по отношению к вам) и негативной (непричинение вреда другим) формах. При этом моральность поведения целиком и полностью определяется схемой сопоставления конспецификов с «Я». Если агент определяет конспецика как своего, то, в силу наличия ассоциативных связей между компонентами знаков семиотической системы, формируются возбуждения, определяющие поведенческую реакцию на состояние конспецифика: помощь, подражание, обучение и т.п. Для «чужого» реакция будет обратной. Например, в работе [16] описываются эксперименты, в которых варьировалась «склонность к симпатии» агента. При этом, воспринимая окружающих как «своих», агенты проявляли то, что называется альтруизмом. Если склонность была отрицательна (все – «чужие»), то агент проявлял к конспецификам агрессивность, воспринимая их как угрозу или источник пищи. Это – достаточно очевидно в силу того, что если конспецифик – свой, т.е. почти «Я», то ни о каком причинении вреда по отношению к «Я»-«своему» речи не идет, а поведение по отношению к «своему» в определенном смысле является зеркальным к «Я».

При таком подходе мы вновь приходим к тому, что основная задача морали – это разрешение внутренних конфликтов в социуме. Итак, сформулируем следующее утверждение:

Утверждение №2. В основе моральности поведения лежит результат сопоставления конспецификов с «Я», т.е. определение степени «свой-чужой».

Поднимемся уровнем выше. Рассмотрим теперь, каким образом соотносятся между собой когнитивная карта этической системы (рис. 2) с нашими представлениями об устройстве системы управления анимата.

4. Схема морального поведения. Основная задача когнитивного уровня $\mathbb{C}$ с точки зрения управления поведением – это изменение значимости тех или иных потребностей, влияние на систему восприятия и оценок, подмена «реальных» сигналов рефлекторного уровня $\mathbb{E}$- $\mathbb{AB}$-$\mathbb{R}$. На том же уровне $\mathbb{C}$ реализуется и система этических правил. При этом роль этической системы в поведении анимата сводится к двум основным функциям – оценочной и объяснительной.

Оценочная функция. Этическая схема определяет значимость тех или иных потребностей анимата. Основная задача – влиять на результаты сопоставления конспецифика с «Я». Именно эта функция определяет, что для агента все люди равны, близки, одинаково ценны. Или же определяет, кто или что является доминантом агента, т.е. высшим авторитетом, объ-

ектом, по отношению к которому осуществляется максимально подражательное или подчиненное поведение. Согласно Утверждению №1, агент-анимат-робот будет определять поведение, исходя из текущей актуальной потребности и оценки ситуации. При этом, естественно, может возникнуть ситуация конфликта. Мы уже рассматривали ситуацию, в которой анимат оказывался перед выбором – отдыхать или работать. Предположим, что, удовлетворяя потребность в удовольствии (см. рис. 3), анимат выполняет действие *«Отдыхать»*. Вместе с тем, остается невыполненным действие *«Работать»*, связанное с тем же получением удовольствия. Эмоционально-потребностная система управления устроена таким образом, что невыполнение некоторой требуемой поведенческой процедуры приведет к возрастанию абсолютного значения эмоции, связанной с этим действием. Наличие возрастающего сигнала такой обратной эмоциональной связи может привести в конечном итоге к ситуации переключения поведения на то, которое не было реализовано, но было актуально необходимым.

Объяснительная функция. Итак, в ситуации противоречия анимат все равно стабилизирует свое поведение, выбрав что-то одно. Проблема, однако, в том, что поведение коллаборативного робота должно быть предельно прозрачным, т.е. объяснимым, причем на уровне простых и наглядных цепочек рассуждений. Для этого представляется целесообразным воспользоваться концепцией семантики возможных миров С. Крипке, но в самом концептуальном виде. Путь в векторе активации поведения Y_B обнаружены противоречия (α, $\neg\alpha$), например, при α = *«отдыхать»*. Тогда агент формирует когнитивную карту, в которой это противоречие отсутствует (если робот решил отдохнуть, то в такой карте не будет сущности *«работать»*). Это – в некотором смысле миры w и ассоциированный с ним мир w^*. После чего отрабатывается обоснование (обратный вывод) выбранного действия. Если вывод успешен, то это означает, что агент сумел обосновать выбранное поведение. В противном случае его поведение не сможет быть оправдано моральной схемой. Итак, мы формулируем:

Утверждение №3. Схема моральных представлений нужна для того, чтобы определить значимость потребностей и особенности системы восприятия, а главное – обосновать (оправдать) выбранное агентом поведение.

Заключение. Итак, в условиях принципиально неустранимых противоречий и неполноты при формализации моральных учений, основная задача этических схем – это быть основой для объяснения постфактум, а вовсе не быть мотивирующим фактором. Коллаборативный робот как моральный агент – это агент, который должен всегда уметь найти оправдание своему поведению и объяснить его своему партнёру – человеку-оператору. При этом обязательными компонентами системы управления таких агентов являются эмоционально-потребностная схема («психологический» уровень) и когнитивная надстройка. В этом отношении представленная работа является контрапунктом множеству исследований в области той же психологии морального выбора.

Отчасти такая позиция может быть оправдана тем, что здесь речь идет не о человеке и даже не о высших млекопитающих (монополия человека в вопросах морали уже давно под сомнением, см., например, работу Ф. де Вааля [17]). Мы говорим о техническом устройстве, для которого необходимо предложить вполне практические, технически реализуемые схемы поведения, и в частности – формализацию схем поведения, которые оцениваются с точки зрения этичности. Будет коллаборативный робот, этичность которого определяется способностью к оправданию своего поведения действительно моральным агентом, или, напротив, такой образ функционирования будет считаться глубоко аморальным – это уже другой вопрос, выходящий за рамки настоящего исследования.

Благодарности. Исследование выполнено при частичной финансовой поддержке РФФИ в рамках научного проекта № 20-07-00770.

Список источников

1. Тарасов В.Б. От объяснимого искусственного интеллекта к «понимающим» когнитивным агентам / В.Б. Тарасов // Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов X-й Международной научно-технической конференции (ИММВ-2021, Коломна, 17-20 мая 2021 г.). – Смоленск: Универсум, 2021. – Т.1. – С. 175-189.
2. Апресян Р.Г. Этика: учебник / Р.Г. Апресян // М.: КНОРУС, 2017. – 356 с.
3. Лефевр В. Алгебра совести / В. Лефевр // М.: Когито-Центр, 2003. – 426 с.
4. Karpov V.E. Can a robot be a moral agent? Lecture Notes in Artificial Intelligence (LNAI), 2020, vol. 12412, pp. 61-70, DOI: 10.1007/978-3-030-59535-7_5.
5. Сушенцова М.С. Утилитаризм И. Бентама и Дж. С. Милля: от добродетели к рациональности / М.С. Сушенцова // Вестник СПбГУ. Экономика, 2017. – Т. 33. – № 1. – С. 17-25.
6. Гусейнов А.А. Этика / А.А. Гусейнов, Р.Г. Апресян // М.: Гардарики, 2000. – 472 с.
7. Кузнецов О.П. Интеллектуализация поддержки управляющих решений и создание интеллектуальных систем / О.П. Кузнецов // Проблемы управления, 2009. – № 3.1. – С. 64-72.
8. Кулинич А.А. Компьютерные системы моделирования когнитивных карт: подходы и методы / А.А. Кулинич // Проблемы управления, 2010. – № 3. – С. 2-16.
9. Девяткин Л.Ю. О подлинно паранепротиворечивых и подлинно параполных многозначных логиках / Л.Ю. Девяткин // Логические исследования, 2019. – Т. 25. – № 2. – С. 26-45.
10. Сидоренко Е.А. Релевантная логика (предпосылки, исчисления, семантика) / Е.А. Сидоренко // М.: ИФРАН, 2000. – 243 с.
11. Карпенко А.С. Логика паранепротиворечивая / А.С. Карпенко // Гуманитарный портал: Концепты. URL: https://gtmarket.ru/concepts/6976 (дата обращения: 22.05.2022).
12. Wilson S.W. Classifier Systems and the Animat Problem. Machine Learning, 1987, vol. 2, pp. 199-228.
13. Симонов П.В. Потребностно-информационная теория эмоций / П.В. Симонов // Вопросы психологии, 1982. – Т. 6. – С. 44-56.
14. Simonov V.P. Thwarted action and need – informational theories of emotions. International Journal of Comparative Psychology, 1991, vol. 5, iss. 2, pp. 103-107.
15. Апресян Р.Г. Генезис золотого правила / Р.Г. Апресян // Вопросы философии, 2013. – № 10. – С. 39-49.
16. Карпов В.Э. К вопросу о моральных аспектах адаптивного поведения искусственных агентов / В.Э. Карпов, П.С. Сорокоумов // Искусственные общества, 2021. – Т. 16. – № 2.
17. Waal de F. The Bonobo and the Atheist: In Search of Humanism among the Primates. W.W. Norton & Company, 2014, 320 p.

Карпов Валерий Эдуардович. *К.т.н., доцент, начальник лаборатории робототехники, НИЦ «Курчатовский институт», AuthorID: 338997, SPIN: 3735-1511, karpov.ve@gmail.com, 123182, Россия, Москва, пл. Академика Курчатова, д. 1*

Королева Мария Николаевна. *К.т.н., доцент кафедры «Компьютерные системы автоматизации производства», МГТУ им. Н.Э. Баумана, AuthorID: 712582, SPIN: 6341-2068, maria.svyatkina@gmail.com, 105005, Россия, Москва, 2-я Бауманская ул., д. 5, стр. 1*

UDC 004.8: 17
DOI:10.38028/ESI.2022.28.4.018

On the issue of formalization collaborative robot ethical behavior

Valery E. Karpov[1], Maria N. Koroleva[2]

[1] National Research Center "Kurchatov Institute"

[2] Bauman Moscow State Technical University,

Russia, Moscow, *maria.svyatkina@gmail.com*

Abstract. The paper considers the issues of creating mechanisms that allow a collaborative robot to be considered a moral agent. It is shown that in the conditions of fundamentally irremovable contradictions and incompleteness, the use of logical schemes and various non-classical logics does not solve the problem of formalizing moral teachings and ethics of behavior. The example shows that the task of moral schemes is not only and not so much to determine or evaluate the agent action (behavior), but to implement the explanatory behavior component. Utilitarianism in the understanding of I. Bentham and J. S. Mill in combination with the axiological system of hedonism rules are considered as ethical theories of the robot behavior. Cognitive maps are

used for qualitative analysis. Three fundamental statements of the moral behavior of a robot are formulate in the article.

Keywords: ethics, collaborative robot, non-classical logics, moral agent, animat, emotional-needs architecture, empathy

Acknowledgements: The reported study was funded with partial financial support by RFBR, project number 20-07-00770.

References

1. Tarassov V.B. Ot ob″yasnimogo iskusstvennogo intellekta k «ponimayushchim» kognitivnym agentam [From Explainable Artificial Intelligence to "Understanding" Cognitive Agents]. Proceedings of the the 10th International Conference on Integrated Models and Soft Computing in Artificial Intelligence (IMSC-2021, Kolomna, Russia, May 17-20, 2021), Smolensk, Universum, 2021, vol.1, pp. 175-189.
2. Apresyan R.G. Etika: uchebnik [Ethics: textbook]. M.: KNORUS, 2017, 356 p.
3. Lefebvre V. Algebra sovesti [Algebra of Conscience]. M: Cogito-Center [Cogito-Centre], 2003, 426 p.
4. Karpov V.E. Can a robot be a moral agent? Lecture Notes in Artificial Intelligence (LNAI), 2020, vol. 12412, pp. 61-70, DOI: 10.1007/978-3-030-59535-7_5.
5. Sushentsova M.S. Utilitarizm I. Bentama i Dzh. S. Millya: ot dobrodeteli k ratsional′nosti [Utilitarianism of J. Bentham and J. S. Mill: from Virtue to Rationality]. Vestnik SPbGU. Ekonomika [Bulletin of St. Petersburg State University. Economy], 2017, vol. 33, iss. 1, pp. 17-25.
6. Guseinov A.A., Apressyan R.G. Etika [Ethics]. M: Gardariki, 2000, 472 p.
7. Kuznetsov O.P. Intellektualizatsiya podderzhki upravlyayushchikh resheniy i sozdaniye intellektual′nykh sistem [Intellectualization of Control Decisions Support and Creation of Intellectual Systems]. Problemy upravleniya [Problems of management], 2009, vol. 3.1, pp. 64-72.
8. Kulinich A.A. Komp′yuternyye sistemy modelirovaniya kognitivnykh kart: podkhody i metody [Computer systems for modeling cognitive maps: approaches and methods]. Problemy upravleniya [Problems of management], 2010, vol. 3, pp. 2-16.
9. Devyatkin L.Yu. O podlinno paraneprotivorechivykh i podlinno parapolnykh mnogoznachnykh logikakh [On genuine paraconsistent and cenuine paracomplete newline many-valued logics]. Logicheskiye issledovaniya [Logical Investigations], 2019, vol. 25, no. 2, pp. 26-45.
10. Sidorenko E.A. Relevantnaya logika (predposylki, ischisleniya, semantika) [Relevant logic (premises, calculus, semantics)]. M.: IFRAN, 2000, 243 p.
11. Karpenko A.S. Logika paraneprotivorechivaya [Paraconsistent logic]. Humanitarian Portal: Concepts. Available at: https://gtmarket.ru/concepts/6976 (accessed: 22.05.2022).
12. Wilson S.W. Classifier Systems and the Animat Problem. Machine Learning, 1987, vol. 2, pp. 199-228.
13. Simonov P.V. Potrebnostno-informatsionnaya teoriya emotsiy [Need-Information Theory of Emotions]. Voprosy Psychologii [Questions of psychology], 1982, vol. 6, pp. 44-56.
14. Simonov V.P. Thwarted action and need – informational theories of emotions. International Journal of Comparative Psychology, 1991, vol. 5, iss. 2, pp. 103-107.
15. Apresyan R.G. Genezis zolotogo pravila [Genesis of the Golden Rule]. Voprosy filosofii [Questions of Philosophy], 2020, vol. 58, no. 2, pp. 109-123.
16. Karpov V.E., Sorokoumov P.S. K voprosu o moral′nykh aspektakh adaptivnogo povedeniya iskusstvennykh agentov [On moral aspects of adaptive behavior of artificial agents]. Iskusstvennyye obshchestva [Artificial Societies], 2021, vol. 16., iss. 2.
17. Waal de F. The Bonobo and the Atheist: In Search of Humanism among the Primates. W.W. Norton & Company, 2014, 320 p.

Valery V. Karpov. *PhD, Head of the Laboratory of Robotics, National Research Center "Kurchatov Institute", AuthorID: 338997, SPIN: 3735-151, karpov.ve@gmail.com, 123182, Russia, Moscow, Akademika Kurchatova Square, 1.*

Maria N. Koroleva. *PhD, Associate Professor of the Computer Systems of Production Automation Department, Bauman Moscow State Technical University, AuthorID: 712582, SPIN: 6341-2068, maria.svyatkina@gmail.com, 105005, Russia, Moscow, 2-nd Baumanskaya, 5, b.1.*

Статья поступила в редакцию 25.07.2022; одобрена после рецензирования 01.11.22; принята к публикации 01.11.2022.

The article was submitted 07/25/2022; approved after reviewing 11/01/2022; accepted for publication 11/01/2022.