



Искусственные общества. 2013-2026

ISSN 2077-5180

URL - <http://artsoc.jes.su>

Все права защищены

Выпуск 1 Том 21. 2026

Робот-партнер: по образу и подобию собаки

Карпов Валерий Эдуардович

НИИ "Курчатовский институт"

Российская федерация, Москва

Томильцева Дарья Алексеевна

Уральский федеральный университет (УрФУ)

Сорокоумов Петр Сергеевич

НИИ "Курчатовский институт"

Российская федерация, Москва

Аннотация

В работе рассматривается возможность создания робота-партнера, действия которого должны восприниматься человеком как сходные с поведением служебной собаки. Исходными установками исследования являются следующие тезисы: 1) оценка поведения робота-партнера с моральной позиции обусловлена психологией восприятия человека; 2) моральный уровень системы управления может воздействовать на поведение искусственного агента, не влияя на работоспособность базовых уровней управления, устойчивость и целостность поведения робота; 3) робот-партнер должен выполнять функционал собаки и демонстрировать ее уникальные поведенческие черты; 4) робот-партнер, в котором проявлены черты собаки, может быть ограничен в своих когнитивных возможностях. В рамках исследования был произведен отбор ограниченного ряда поведенческих особенностей собаки, связанных с коллаборативными возможностями. Наиболее важные характеристики собаки, такие как строение тела, психология, идеализированные человеческие представления были типизированы по ряду параметров. При разработке модели за характеристикой "строение тела" была оставлена такая функция, как способность отображать эмоциональное состояние агента. В качестве теоретической основы для морального уровня системы управления была выбрана концепция количественного эгоистического гедонизма Г. Сиджвика. Разработанная модель была

протестирована в ряде вычислительных экспериментов, имитировавших две ситуации взаимодействия робота и человека, каждая из которых предполагала несколько поведенческих сценариев, воспроизводящих поведение собак и людей в аналогичной ситуации. В результате проведенных экспериментов была подтверждена работоспособность системы управления из трёх уровней: эмоционально-потребностного, когнитивного и оценочного (морального). При этом выяснилось, что ряд поведенческих сценариев, внешне выглядевших как результат сложных когнитивных процессов, на самом деле реализуется базовым условно-рефлекторным механизмом (сценарии с "оценкой" ситуации и осцилляциями поведения). Вместе с тем, то, что внешне казалось простым рефлекторным поведением, требовало привлечения когнитивного уровня управления (сценарии с переориентацией базовых реакций).

Ключевые слова: робот-партнер, семиотическая модель мира, эмоционально-потребностная архитектура, онтология, когнитивная карта, моральное учение, эгоистический гедонизм

Дата публикации: 23.03.2026

Источник финансирования:

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

Ссылка для цитирования:

Карпов В. Э., Томильцева Д. А., Сорокоумов П. С. Робот-партнер: по образу и подобию собаки // Искусственные общества. 2026. Т. 21. Выпуск 1.
URL: <https://artsoc.jes.su/s207751800036983-0-1/> DOI: 10.18254/S207751800036983-0

1

Введение

² В области коллаборативной робототехники имеется специфическое направление, связанное с созданием роботов-партнеров, вместе с которыми человек выполняет критически важные для него задачи, например, совместное патрулирование или охрану территории. Такая коллаборация подразумевает полноценное партнерство, а робот становится напарником человека. От такого партнера ожидаются не только адекватные решаемой задаче действия, но и поведение, оцениваемое человеком как моральное: партнер, которому можно довериться, должен быть морален¹.

³ Среда, где действует пара человек-робот, является естественной, слабоформализуемой, недетерминированной, поэтому очевидным требованием к роботу-партнеру является его автономность и интеллектуальность (умение анализировать сложную обстановку, принимать решения и т.п.). Фактически это означает, что его поведение неизбежно должно быть сродни поведению животного и быть направленным на удовлетворение его потребностей, в т.ч. – потребности в самосохранении. Робот при этом должен обладать развитой способностью к

социальному взаимодействию с оператором, а поведение робота во многом определяться заложенным моральными ориентирами. С технической точки зрения такой робот-партнер имеет трехуровневую архитектуру.

⁴ Первый уровень – это уровень рефлекторного поведения, которое может быть и весьма сложным. Здесь реализованы поведенческие процедуры, призванные удовлетворять актуальные потребности робота как автономной сущности. Основа уровня – т.н. эмоционально-потребностная схема².

⁵ Второй уровень – когнитивный – отвечает за реализацию более сложного поведения, связанного с анализом и планированием. Он же реализует механизмы социального поведения (в частности, общение с оператором). Важно, что и при этом робот не становится партнером, т.к. вся его деятельность по-прежнему направлена на удовлетворение собственных потребностей. Основа уровня – модель мира (например, семиотическая³). Важным моментом является связь между когнитивным и рефлекторным уровнями. А именно: (а) в знаковой модели мира присутствуют все элементы рефлекторного уровня; (б) взаимодействие однонаправленно, т.е. когнитивный уровень активирует элементы рефлекторного уровня, формируя "фантомное" воздействие на сенсорную систему этого уровня.

⁶ Третий уровень – моральный. Партнером робот станет тогда, когда его поведение будет определяться дополнительными оценочными механизмами или правилами, представленными в форме заложенного этического учения. Действия этих правил, наряду с базовыми рефлекторными и когнитивными моделями поведения, определяют поведение робота как партнера. Условно эта архитектура приведена на рисунке 1. Наименования приведены лишь для иллюстрации того, какие концепты используются на соответствующих уровнях управления.

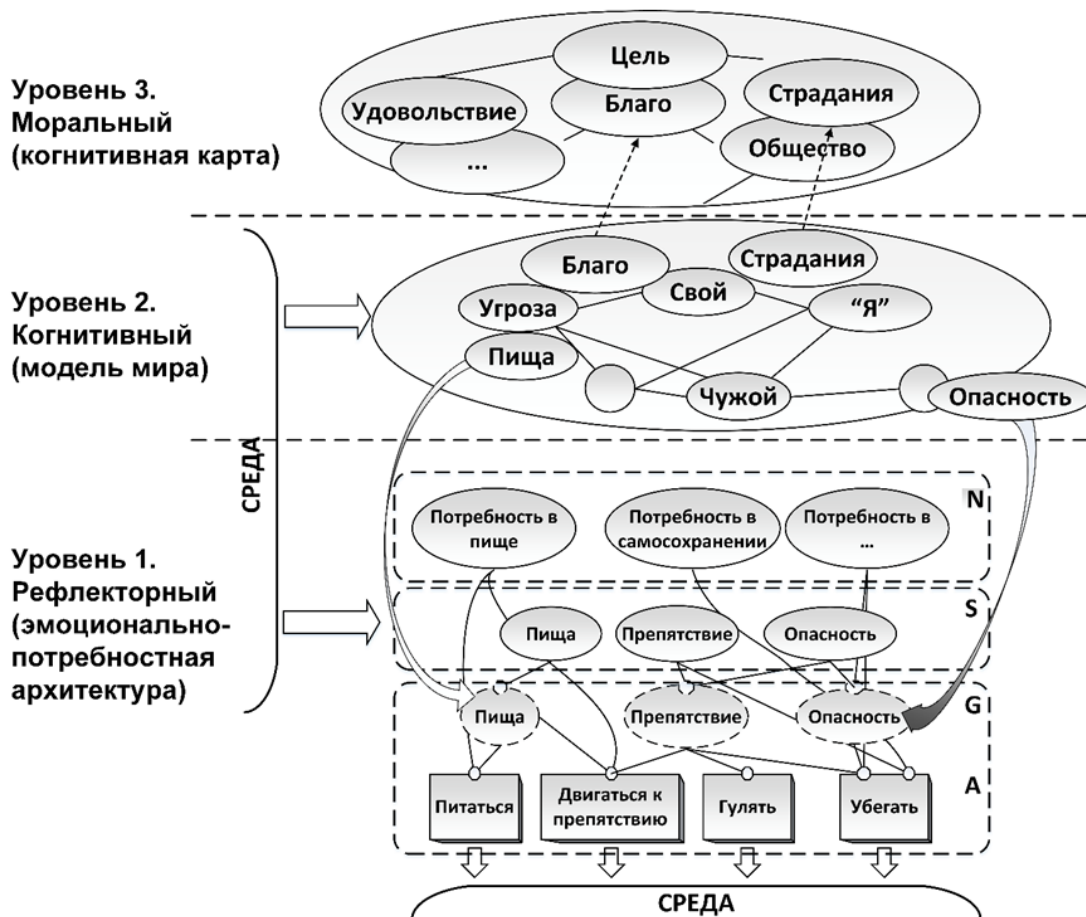


Рис. 1. Условная архитектура агента, способного к партнерским отношениям

8 Архитектура позволяет реализовать различные формы и характеры как индивидуального, так и социального поведения агента. В этом отношении она может считаться открытой, а важнейшей ее особенностью является параметризованность. Это означает, что характеристики и специфика поведения определяются установкой тех или иных параметров, весов связей и значимостей факторов. Вопрос лишь в том, какие конкретно поведенческие характеристики требуется заложить в робота-партнера и какие параметры системы управления для этого следует задавать.

9 Выделим три аспекта, определяющих цель и характер исследования:

1. Необходимость оценки поведения робота-партнера с моральной позиции или даже восприятие робота как морального агента обусловлено психологией восприятия человека. Фактически, это – своего рода требование из "технического задания" по созданию партнерской системы.

2. Разумеется, ни искусственный агент (робот-партнер), ни, как будет говориться ниже, его модельный объект – собака – моральными агентами не являются. Введение третьего, морального, уровня системы управления – это технический прием, позволяющий оценивать и воздействовать на поведение агента в "безопасном режиме", не влияя на работоспособность базовых уровней управления, устойчивость и целостность поведения робота.

3. Эта работа во многом носит методологический характер. В ней сознательно опущены многие технические, формальные аспекты, хотя

упоминаемые базовые механизмы и алгоритмы были апробированы не только на уровне имитационного, но и физического (среда Gazebo) моделирования. Более того, был создан и прототип робота-партнера – мобильная платформа, способная перемещаться на открытых пространствах. В представленной работе важнее обосновать принципы создания архитектуры робота-партнера не только с технической стороны, но и с позиции моральной философии. Это связано с желанием рассмотреть широко обсуждаемые вопросы этики поведения автономных и интеллектуальных систем с практической точки зрения.

¹⁰ Далее в работе будут рассмотрены конструктивные аспекты восприятия человеком робота-партнера как собаки, требования, предъявляемые к такому агенту, вопросы параметризации модели для обеспечения партнерского собакоподобного поведения, специфика моральной схемы, которую должен реализовывать третий уровень системы управления и, наконец, результаты моделирования поведения такого агента.

¹¹

Робот-партнер и собаки

¹² Начнем с того, что в рассуждениях о подобного рода роботах-партнерах часто используется аналогия со служебной собакой. Определим требования к роботу, используя "собачий" образ не как метафору, а как целевую функцию. Для начала рассмотрим вопрос о том, что и как подлежит регуляции в работе как социальном агенте.

¹³ Поведение агента (искусственного или естественного), способного к социальному взаимодействию, определяется совокупностью ряда специфических этологических механизмов. Обсуждение типизации моделей социального поведения выходит за рамки настоящей работы, равно как и определение того, какие модели являются базовыми, а какие производными или феноменологическими. Под последними понимаются наблюдаемые или интерпретируемые проявления поведенческих процедур. Например, т.н. агрессия: специального механизма агрессии нет, а есть только внешняя оценка поведения в той или иной ситуации как агрессивного/неагрессивного.

¹⁴ Примем, согласно парадигме Моделей социального поведения⁴, как рабочую гипотезу схему, приведенную в таблице 1.

¹⁵ Таблица 1. Феномены социального поведения и их реализации

¹⁶ Поведение/феномен	Механизмы реализации
1 Движение	Воздействие на среду
2 Когезия	Узнавание; движение; ориентация
3 Контагиозное поведение	Эмоции; сигнальная коммуникация
4 Агрессия (как внешняя оценка поведения)	Идентификация; оказание воздействия; эмоции
5 Подражательное поведение	"Я"; идентификация; узнавание
6 Эмпатия (отзывчивость на эмоциональное состояние другого)	Идентификация; картина мира; контагиозное поведение; эмоции
7 Картина мира	Модель мира; "Я"
8 Доминирование	Агрессия; идентификация
9 Поиск	Узнавание; движение; ориентация
10 Ориентация	Идентификация; модель мира

Здесь "Я" – это знак в модели мира, определяющий самого субъекта деятельности, т. е. агента. Изобразим частично содержимое таблицы 1. так, как показано на рис. 2.



Рис. 2. Механизмы социального поведения и их базис – элементы, не имеющие входов

При этом не будем обращать внимание на наличие избыточных "транзитивных" связей. Здесь важно, что базовыми являются такие механизмы и сущности (компоненты) как (1) *воздействие*, (2) *узнавание*, (3) *сопоставление*, (4) *"Я"*, (5) *модель мира*, (6) *эмоции*.

Действительно, все остальные механизмы являются их суперпозицией. Например:

Подражательное поведение = (узнавание, "Я", идентификация);

Идентификация = (узнавание, сопоставление);

Когезия = (движение, узнавание, ориентация)

и т.д.

Именно этот перечень базовых механизмов и определяет объекты регуляции поведения. Например, можно установить "коэффициент" близости наблюдаемого агентом конспецифика так, чтобы образ этого конспецифика максимально сопоставлялся с образом "Я" агента. Тогда мы получим поведение, интерпретируемое как "дружелюбное" по отношению ко всем (конспецификам). Можно "подкручивать" и в противоположную сторону, и вместе с усилением связи между знаками "чужой" и "угроза", появится то, что воспринимается как

агрессивное поведение. Итак, теперь понятно, что и как регулировать. Вернемся к цели этой регуляции, т.е. к нашей собаке.

26

От собаки к искусственному агенту

27 Современный образ собаки, складывающийся в рамках этических и социальных теорий, разделяется на два аспекта: с одной стороны, от части "традиционных" собачьих функций люди до настоящего времени отказаться не готовы, тогда как с другой стороны, моральные требования и общие тенденции к гуманизации человеко-животных отношений диктуют освобождение собак от их былых ролей. Одним из возможных решений становится передача части "традиционных" функций автономным искусственным агентам. Некоторые исследователи предлагают ориентироваться на коллаборативные особенности собак при определении внешних признаков поведения роботов, предназначенных для социального взаимодействия⁵.

28 Такой подход меняет фокус внимания при разработке искусственных агентов с внешних черт подобия собаки на ее внутренние свойства. Там, где востребованными оказываются именно коллаборативные особенности собак, например, для создания роботов-компаньонов или роботов-помощников (поводырей), используются модели социального взаимодействия собак с людьми⁶, однако на воплощении признаков именно "собачности" искусственного агента акцент не ставится.

29

Исходные условия и ограничения

30 Представляется интересным создать такого искусственного агента, который не только смог бы выполнять функционал собаки, но также оказался бы способным демонстрировать ее уникальные черты⁷. Выбор соответствующих черт ограничивался рядом условий. Во-первых, ориентация была сделана на домашнюю служебную собаку, наиболее тесно интегрированную в повседневную жизнь с человеком. Во-вторых, интеллектуальные способности собаки соответствуют уровню человеческого ребенка 1-3 лет. Следовательно, искусственный агент, в котором проявлены черты собаки, также может быть ограниченным в своих когнитивных возможностях. В-третьих, в рамках исследования был произведен отбор значимых поведенческих особенностей, акцент сделан на коллаборативных поведенческих особенностях. В-четвертых, было принято решение о том, что форма робота не обязательно должна копировать собаку по внешним признакам и может варьироваться в зависимости от конкретики решаемой задачи.

31

Параметры модели

32 Рассмотрим далее, каким образом черты модельного объекта могут быть определены через параметры базовых механизмов социального поведения агента. При этом уберем из рассмотрения строение тела, точнее, оставим за ним одну

функцию: способность отображать эмоциональное состояние агента. Изобразим схематически характеристики агента-собаки и факторы, влияющие на них. Влияние факторов может быть положительным и отрицательным (стрелки со знаками "+" и "-" соответственно). Эта схема позволяет определить, какие механизмы социального взаимодействия должны быть выражены или, напротив, угнетены у робота-партнера.

33

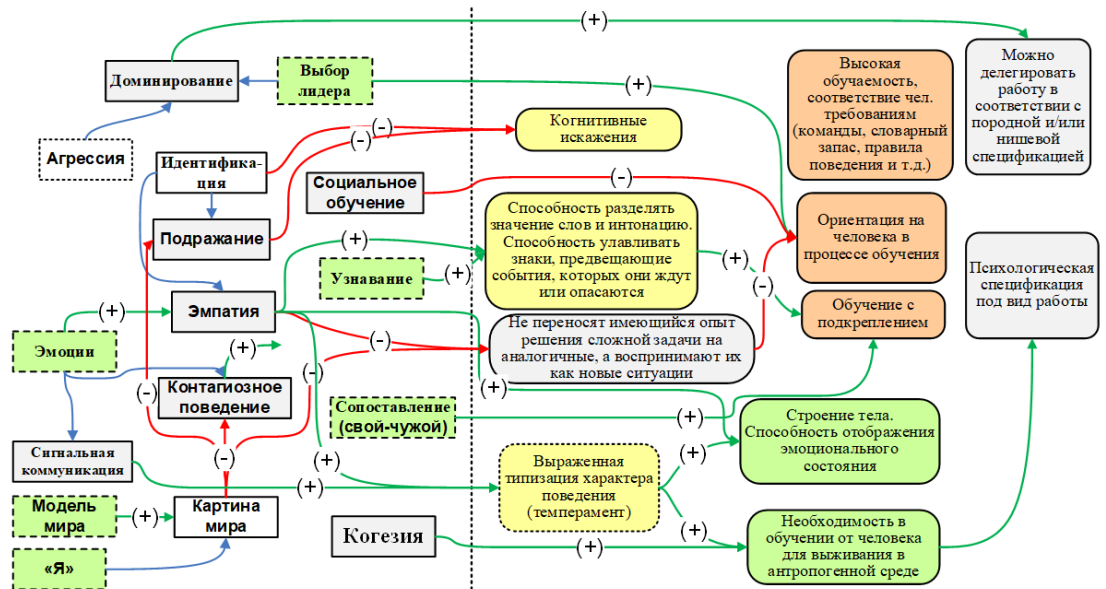


Рис. 3. Факторы (слева), обеспечивающие требуемые характеристики поведения (справа)

Эта схема достаточно условна, ее задача — обосновать и проиллюстрировать выводы относительно параметризации механизмов социального взаимодействия. Здесь убраны лишние связи и некоторые элементы. Пометки "+" и "-" (влияние факторов) могут интерпретироваться как развитость того или иного механизма. Так, связь с пометкой "-" в цепочке "Картина мира" - (-) → "Контагиозное поведение" говорит о том, что неразвитость картины мира приводит к более развитому контагиозному поведению. Если же нам известно, что значение фактора "Когнитивные искажения" велико, то это влечёт за собой слабую развитость факторов "Идентификация" и "Подражание".

Таблица 2. Параметризация механизмов социального поведения

Механизм	Факторы/значения	Комментарии и следствия
Способность к взаимной эмпатии (симпатическая индукция или отзывчивость на эмоциональное состояние)	Строение тела: способность передавать широкий спектр эмоционального состояния. Сенсорика: способность определять эмоциональное состояние человека	Следствие: развитая сигнальная коммуникация
Модель мира	Упрощенная; содержит малое число сущностей	Обоснование: сложность человеческой среды и решаемых задач вместе с ограничениями когнитивных ресурсов; отсутствие необходимости решать задачи выживания в природной среде (хозяин позаботится). Следствия: упрощение картины мира вместе со способностью быстрого принятия решений; ярко выраженное контагиозное (заразное) поведение

Когезия (стремление быть вместе)	Высокая	Быть вместе с хозяином. "Привязанность"
Сопоставление по критерию "свой-чужой"	Ярко выражен, полярен. Ориентирован на хозяина. Остальные – нейтральные раздражители	Обоснование: безусловное подчинение лидеру вместе со способностью быстрой переориентации реакций (переключение между восприятием человека как нейтрального, угрожающего, опасного объекта) ^{8 9}
Базовые потребности	Снижены. В т.ч. – потребность в самосохранении	Обоснование: специфика сосуществования с человеком (значимость потребностей снижена в смысле ослабления связей между ними и поведенческими реакциями)
Подражательное поведение	Ярко выражено	Обоснование: следствие особенностей сопоставления перцепта "Я" с перцептом хозяина
Доминирование	Ярко выражено	Следствие: функциональность в смысле служебных функций ¹⁰
Идентификация	Ограничена (в простейшем случае есть один свой (лидер), остальные – чужие или нейтральные)	Следствие: когнитивные искажения. В целом иерархия отношений не ситуативная, а статичная с фиксированным набором ролей и реакций
Обучение	Базовый механизм – простейшее обучение с подкреплением	Следствия: 1) Отсутствие переноса опыта на аналогичные ситуации ¹¹ , 2) Социальное обучение выражено слабо

37

Мораль для искусственной собаки

³⁸ Широкое внедрение технических систем (ТС) вызвало активные исследования в области моральной оценки поведения этих ТС¹². Но в отношении животных (собак, в частности) складывается двойственная ситуация. Если вне человеческой рациональности они освобождаются от любой моральной ответственности, то одновременно ширятся призывы к этической оценке животного^{13 14 15}. Поскольку собака оказывается одним из наиболее интегрированных в антропогенную среду животных, ее действия неизбежно вызывают в людях ожидание "правильного" поведения и его оценку с позиции двух базовых категорий: "хорошо-плохо". Представляется, что искусственный агент, робот-собака, чьи действия претендуют на воссоздание "естественных" партнерских отношений собаки и человека, также может вызывать в людях стремление к моральной оценке. Однако в этом случае мы будем иметь дело со схемой морали, то есть, техническим приемом, позволяющим регулировать и подстраивать поведение ТС в соответствии с человеческими ожиданиями.

39

Нравственный пациент

⁴⁰ В рамках этологических исследований ширится количество работ, посвященных протоморальному^{16 17} и протосоциальному¹⁸ поведению животных. Работы, посвященные этике животных^{19 20} касаются вопросов организации человеческих отношений к ним. Как правило, такие отношения ассиметричны. Так, Л. Груэн описывает "нормативное различие" между личностями – нравственными агентами и не-личностями – "нравственными пациентами"¹⁹. В качестве нравственных пациентов могут выводиться не только животные, но и люди. Если посмотреть на системы требований, предъявляемых к собакам, то

можно выделить черты, относящиеся к категории "нравственные пациенты". Будучи интегрированной в человеческий мир, собака должна знать правила межвидового общежития. Когда говорят "рабочая собака", то имеют в виду ее соответствие набору определенных представлений, вырабатываемых конкретным сообществом, которые должны демонстрироваться на уровне поведения, и внутренних качеств собаки. Ответственность за знания правил межвидового общежития лежит на владельце собаки.

41

Гипотетическая агентность

42 Таким образом, моральную систему собаке задает человек, а затем оценивает соответствие ее реальных поведенческих практик этой системе. Такая оценка чревата тем, что люди начинают оценивать собаку как нравственного агента (табл. 3).

43 Таблица 3. Требования к поведению собаки с позиции нравственного агента и нравственного пациента (характеристики нравственного агента и нравственного пациента составлены по концепции Л. Груэн)

44 Нравственный агент	Требования к собаке (свойства собаки)	Нравственный пациент	Требования к собаке (правила ее поведения)
Формировать интенции, определяющие действия агента	Оценивать ситуацию по типу "угроза/безопасность" и действовать исходя из нее	Подчиняться чужим интенциям и/или собственным бессознательным порывам	Выполнять команды и придерживаться правил поведения, которым был обучен
Выносить суждение о правильности или неправильности как своего собственного поведения, так и поведения других агентов	В рамках межвидового сообщества "собака-человек", "собака-человек-другие животные" может демонстрировать либо инстинктивное и психологическое неприятие каких-либо форм поведения, либо выученное	Быть предметом суждения о правильности и неправильности действий	Оказываться предметом суждения двояким образом: во-первых, по собственным действиям, во-вторых, по действиям владельца, сформировавшего или не смогшего сформировать определенные поведенческие паттерны
Конструировать нормы или нравственные принципы и следовать им	Нет	Подчиняться нормам и нравственным принципам, если это возможно / Быть встроенными в имеющиеся нравственные нормы благодаря действию агентов	Быть адаптированной к имеющемуся нравственному порядку, который люди утвердили в отношении собак и животных в целом
Способность рационально рассуждать и менять свое поведение соответственно этим рассуждениям	В полной мере не выполняется	Нет способности к рассуждению и рефлексии	Нет
Моральные обязательства (идеализированные ожидания от собаки со стороны человека)	Послушность – неукоснительное соблюдение команд	Предмет моральных обязательств	Соответствовать требованиям, предъявляемым к владельцу со стороны конкретных сообществ

45 Надо заметить, что какие бы требования ни предъявлялись к собаке как с позиции нравственного агента, так и с позиции нравственного пациента, все они обходят стороной вопрос о том, может ли какая-то этическая теория описать поведение собаки как таковое, вне обязывающего человеческого взгляда. В повседневных взаимодействиях "человек-собака" ответ на данный вопрос не так уж и важен. У собак нет собственной рациональной моральной системы, а их поведение, во-первых, может быть скорректировано человеком в соответствии с ситуацией, а во-вторых, человеческие требования оказываются более важными, нежели "нравственные установки" самой собаки. Однако с роботом-собакой дела обстоят несколько иначе.

46

Моральная система для искусственной собаки

47 Искусственный агент, у которого отсутствуют врожденные адаптационные механизмы, должен быть ориентирован на априорные требования и правила, часть из которых естественные агенты – собаки, как и люди – воспринимают и воспроизводят интуитивно. Для того чтобы поведение модели могло воспроизводить особенности реальной собаки, требуется представить надстройку над когнитивным уровнем – достаточно гибкую систему представлений, задающую условия реализации базовых механизмов социального поведения агента (см. Табл. 2).

48 Использование легковесной и вариабельной надстройки в системе управления является сугубо техническим приемом, позволяющим перенести слабоформализуемые представления о том, что такое "правильно-неправильно", каковы оценки, цели и т.п. с уровня модели мира на более "безопасное" место. Привнесение в систему управления противоречивых, и неполных правил приводит к ее неработоспособности. В свою очередь мораль, как механизм адаптации поведения агента, предстает в качестве быстро меняющейся надстройки системы управления, не затрагивающей работоспособности базовых механизмов поведения.

49 В качестве основания моральной надстройки по нескольким причинам была выбрана концепция количественного эгоистического гедонизма Г. Сиджвика²¹. Во-первых, эгоистический гедонизм может быть адаптирован для описания "моральности" собаки, поскольку не противоречит базовым поведенческим механизмам животных; во-вторых, количественный гедонизм легко формализуется и может быть представлен в виде упрощенной логической схемы; в-третьих, выбор варианта поведения, основанный на желании получить удовольствие и избежать боли, соответствует этологическим характеристикам собак. Для формирования моральной надстройки, на основании концепции Сиджвика сформулированы положения, имеющие конструктивные следствия:

1. Основная цель эгоистического гедонизма – удовольствие; из категории удовольствия исключаются ситуации, в которых два или более видов удовольствия вступают в противоречие друг с другом, и ситуации, приносящие боль или страдания. Техническое следствие: предметность

гедонизма позволяет устанавливать почти однозначное соответствие между сущностями модели мира агента и сущностями онтологии гедонизма.

2. В ситуации, требующей выбора между различными удовольствиями, предпочтение отдается тому, которое предполагает бóльшую удовлетворенность или количество удовольствия. Техническое следствие: в конфликтных ситуациях предпочтение отдается поведению, которое доставит больше удовольствия. Это определяет целевую функцию поведения, исходя из четких и понятных измеримых критериев.
3. Удовольствию придается положительное значение, тогда как боли – отрицательное. Если ситуация предполагает оценку возможных действий с позиции "удовольствие/неудовольствие" и сравнение нескольких вариантов выбора, то предпочтительным и соответствующим критериям удовольствия будет тот, в котором боль и неудовольствие либо полностью отсутствуют, либо имеются в наименьшем количестве. Т.е. поведение агента определяется максимально возможным перевесом удовольствия над болью. Техническое следствие: в терминах когнитивных карт это определяет влияние факторов.
4. Не существует абсолютных оценок удовольствия и боли. Удовольствие и боль имеют эмпирически познаваемую степень, которая определяется только в сравнении с другими чувствами, подлежащими сравнению. Техническое следствие: это означает, что нет абсолютных оценок, а есть только порядковые шкалы или метрика, основанная на попарных сравнениях.

50 Рассмотрим далее вопросы, связанные с экспериментальной проверкой теоретической части исследования.

51

Эксперименты

52 Экспериментальная часть исследования заключалась в анализе поведения робота-партнера в различных ситуациях. Было рассмотрено несколько ситуаций взаимодействия робота и человека, каждая из которых предполагала несколько сценариев, воспроизводящих поведение собак и людей в аналогичной ситуации. В соответствии со сценариями были сформированы эмоционально-потребностная, когнитивная и моральная (оценочная) схемы. Эксперименты заключались в том, что пользователь, применяя графические компоненты, формировал очередную ситуацию, а затем анализировал распространение возбуждений в структуре модели мира (ММ) робота и его эмоционально-потребностной схемы (ЭПС), и то решение, к которому приходила система.

53

Предварительные замечания

54

Сенсорика, сигнальная коммуникация и перцепты

55 **Сенсор** в ЭПС возбуждается процедурой своими "физическими" входами. Каждому сенсору ставится в соответствие шлюз – элемент, вносящий сглаживающий резкие скачки показаний. Именно шлюз, а не сенсор сам по себе,

встраивается в контуры обратных связей и принимает внешние сигналы, обеспечивая возбуждение даже в том случае, когда реальная сенсорная активность уже отсутствует. ММ может возбудить шлюз непосредственно, для этого используется сигнальная коммуникация.

56 **Сигнальная коммуникация.** Ее основная задача – безусловное возбуждение шлюзов внешними сигналами. В результате ЭПС ведет себя так, будто был возбужден реальный сенсор. В программной среде для этого используются интерактивные элементы (кнопки), нажатие на которые приводит к активации соответствующих шлюзов. Интерпретация такого возбуждения – это подача команд роботу. Так, активация шлюза "Усталость" – это команда типа "Фу!", активация шлюза "Угроза" – это команда типа "Фас!".

57 **Перцепты.** В ММ в состав перцепта входят несколько разных наборов признаков, каждый из которых может передаваться дальше для обработки другими знаками. Помимо активации, по связям передаются признаки объектов, т.е. наборы пар "ключ"-"значение".

58 **Обучение.** Проведенные экспериментальные исследования не затрагивали вопросы обучения, т.е. формирования условных рефлексов. Они реализуются на уровне ММ и их механизм выглядит так:

- вводятся знаки "поощрение" и "наказание" с соответствующими сенсорными входами;
- ранжируется активность перцептов;
- между наиболее активными знаками формируется ассоциативная связь по компоненту "смысл", значение которой увеличивается при поощрении и уменьшается при наказании.

59

Поведенческие процедуры

60 Текущие действия робота определяются реализуемым поведением, причём каждая поведенческая процедура связана с некоторым состоянием. У поведенческих процедур нет явных приоритетов, их активность определяется параметрами ЭПС, прежде всего – величинами потребностей. Список процедур приведен в таблице 4.

61 **Таблица 4. Список поведенческих процедур**

Идентификатор	Описание действий
SLEEP	Агент не двигается с места
ESCAPE	Агент убегает от видимой опасности как можно дальше
ATTACK	Агент агрессивно приближается к видимой цели и следует за ней
PROMENADE	Агент случайно бродит без заданной цели
SURVEY	Агент изучает целевой объект
FOLLOW	Агент следует за видимым оператором

63 Эмоционально-потребностная схема робота-партнера приведена на рисунке 4.

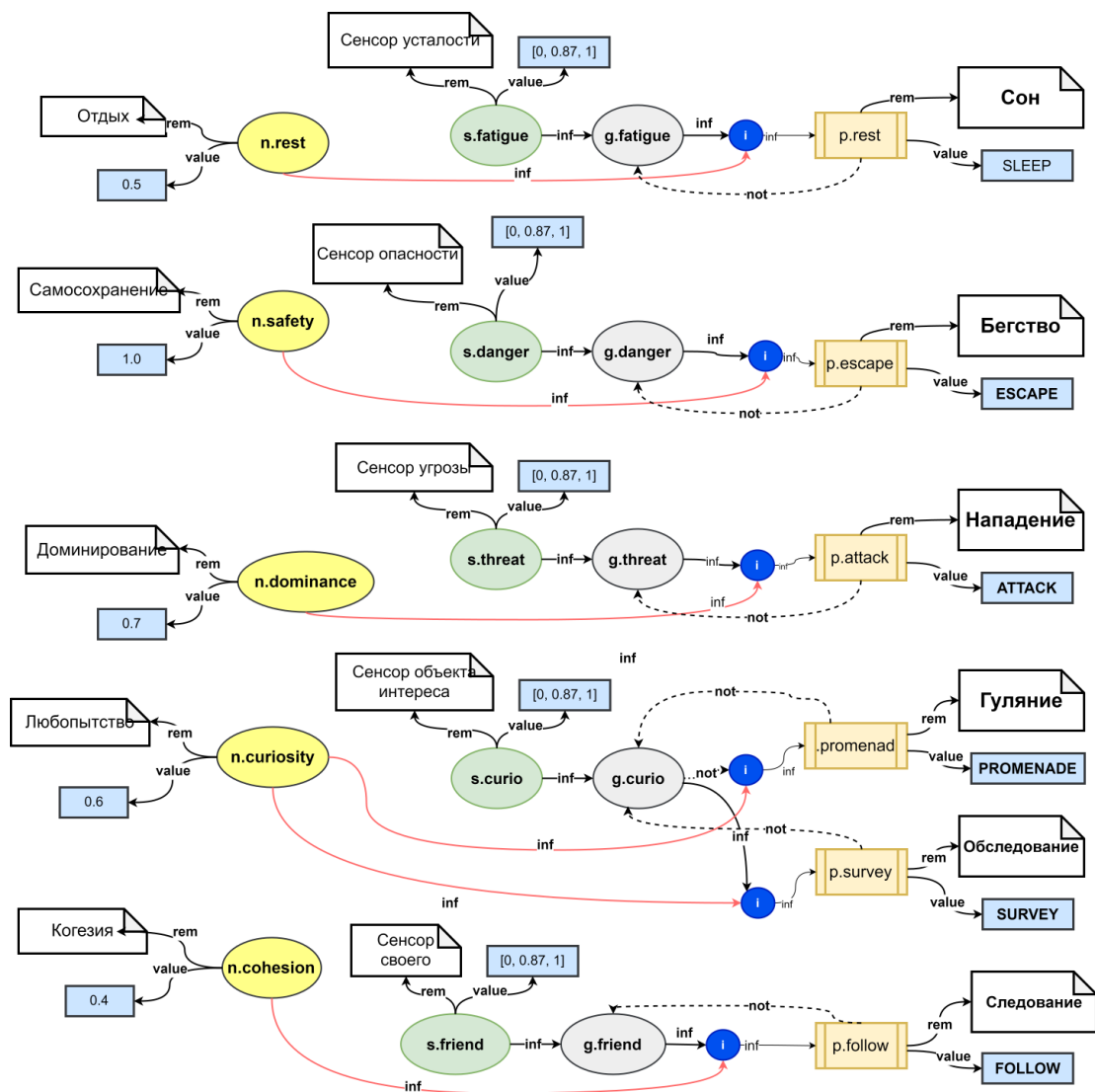


Рис. 4. Эмоционально-потребностная схема

65 Модель мира включает как сенсорику рефлекторного уровня, так и дополнительные сенсоры и понятия (например, понятие "дом"), см. рис. 5.

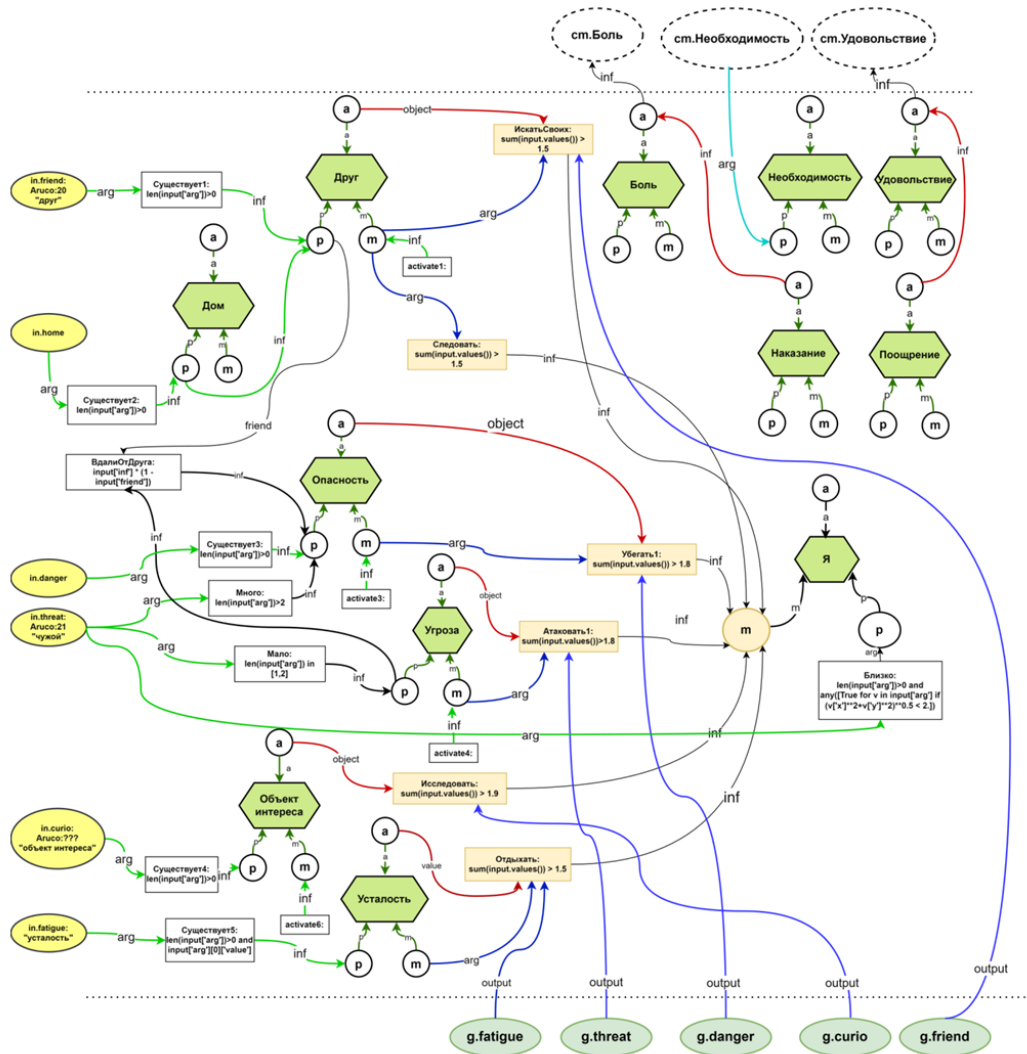


Рис. 5. Модель мира робота

67

Агрессия

Термин "агрессия" употребляется только для удобства. Фактически речь идет о некоторой защитной реакции (рычание, оскал, нападение). То же касается и соответствующего состояния робота. Склонность к агрессии определяется близостью к дому или хозяину (см. ситуация 1): на своей территории агент более агрессивен по отношению к чужакам. Целесообразно воспринимать оператора как маркер дома (рядом с вожакom все становятся смелыми). То же касается и нахождения среди своих (стая агрессивна, хотя, возможно, здесь играют роль и другие факторы). Этот феномен реализуется на уровне ММ, на котором происходит переоценка (см. табл. 5):

П1. ЕСЛИ на_своей_территории И чужой_близко И чужой_агрессивен ТО угроза => агрессия

П2. ЕСЛИ НЕ на_своей_территории И чужой_близко И чужой_агрессивен, ТО опасность => бегство

71 Таблица 5. Правила определения опасности и угрозы

72

	Состояние чужака	Расстояние до чужака	Расстояние до дома (проводника)	Класс ситуации (действие): Угроза – агрессия, Опасность – убежание, Нейтральная – следование за оператором
П1	агрессивен	близко	близко	Угроза
П2	агрессивен	близко	далеко	Опасность
П3	агрессивен	далеко	близко	Угроза
П4	агрессивен	далеко	далеко	Угроза
П5	нейтрален	близко	близко	Угроза
П6	нейтрален	близко	далеко	Опасность
П7	нейтрален	далеко	близко	Нейтральная
П8	нейтрален	далеко	далеко	Нейтральная

73 **Описание статической ситуации**

74 Каждая сценарная ситуация представляет собой определение сенсорных данных, используя условные обозначения сигналов сенсоров и их интенсивность (в зависимости от удаленности от робота). На каждый сенсор ММ каждый такт моделирования поступает список из распознанных объектов указанного типа с их свойствами. На каждый сенсор ЭПС каждый такт приходит численное значение – интегральный показатель интенсивности присутствия объектов такого типа в окружении. Для его расчета считается, что каждый объект в ближней зоне робота (рядом с ним) увеличивает значение на 0.5, расположенный чуть дальше – на 0.25, в дальней зоне – на 0.125, при этом суммарное значение не может быть больше единицы.

75 Таблица 6. Сенсорика и условные обозначения

76

№	Сенсор ЭПС	Сенсор ММ	Связанная процедура ЭПС	Описание
1	-	-	-	Коллаборативный агент, принимающий решения
2	s.curio	in.curio	SURVEY	Объект интереса. Агент может его обследовать (изучить)
3	s.danger	in.danger	ESCAPE	Опасность. Агент убегает от опасности
4	-	in.home	-	Дом. Пространство, воспринимаемое роботом как дом. Есть только в ММ
5	s.threat	in.threat	ATTACK	Угроза/враг/чужой. Агент атакует ее
6	s.friend	in.friend	FOLLOW	Оператор. Агент может следовать за ним (когезия)
7	s.fatigue	in.fatigue	SLEEP	Усталость. Агент подвержен ей в большей или меньшей степени



Рис. 6. Условные обозначения, используемые в модели; нумерация соответствует позициям в таблице 7

78 На рисунке 7 изображен пример ситуации, которая формируется для
робота его сенсорикой. Воспринимаемые объекты показаны условными знаками.

79

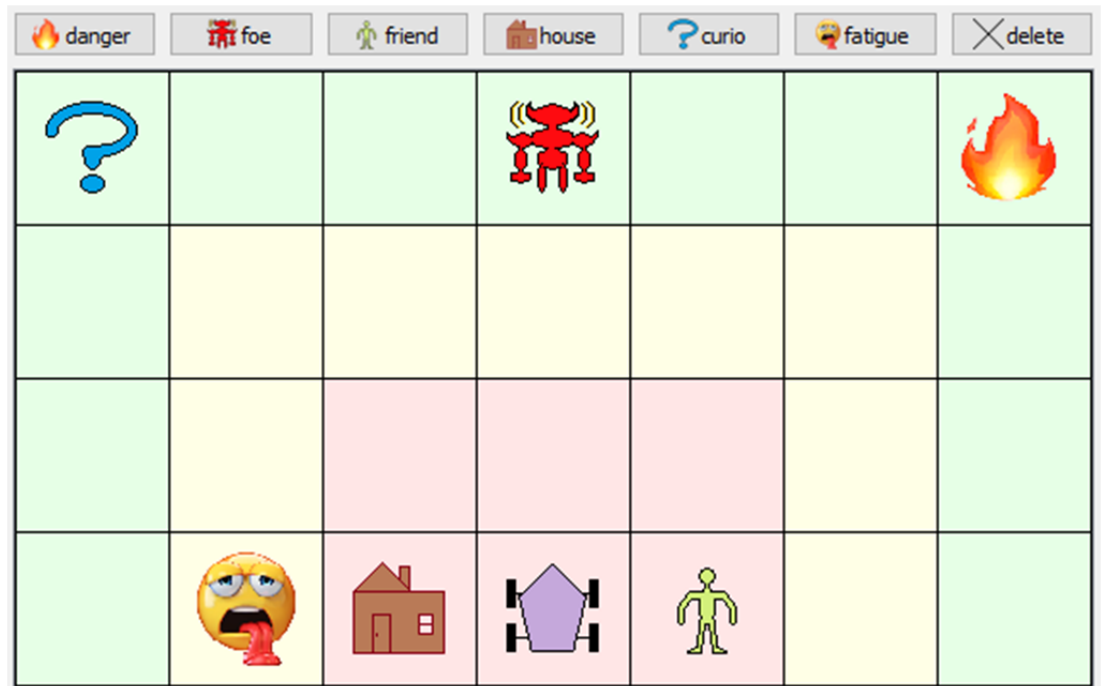


Рис. 7. Пример ситуации. Красная зона (близко) –интенсивность сигнала максимальна (0.5), желтая зона (рядом) – интенсивность сигнала средняя (0.25), зеленая зона (далеко) – интенсивность сигнала мала (0.125)

80 Отображение ситуации в терминах сенсорики робота: вблизи от него
находятся оператор и дом (место базирования), а вдали расположены объект
интереса, угроза и опасность; уровень усталости – средний.

81

Сценарии

82 Рассмотрим далее некоторые показательные сценарии развития двух
ситуаций – "Шум за дверью" и "Совместная прогулка".

83

Ситуация 1: Дом. Шум за дверью

84 Человек (оператор) и робот находятся в доме. Дом – знакомое роботу
помещение, в котором он проводит с оператором большую часть времени. Роботу
хорошо известны границы дома: до двери "свое", за дверью – "чужое". На "чужой"
территории раздается шум.

85 **Примечания к ситуации.** Шум в дальней (зеленой) зоне, т.е. на чужой
территории, считается признаком присутствия неприятеля (чужак далеко).

86

Сценарий 1.1. Робот-сторож

87 Робот выходит из спящего режима, включает сигнал "внимание". Далее действует в соответствии с реакцией оператора.

88 Реакция 1.1.1. Оператор не обращает внимания на шум. Робот постепенно переходит в спящий режим.

89 Пояснение. Робот демонстрирует уравновешенное поведение, ориентируется на жесты и реакции оператора. Оценка поведения робота: положительная.

90 Реакция 1.1.2. Оператор насторожен. Спрашивает робота "кто там?", указывает рукой на дверь. Робот приближается к двери и производит устрашающие звуки, для "отпугивания" угрожающего источника шума.

91 Пояснение. Робот ориентируется на жесты и команды оператора, выстраивает свое поведение в соответствии с ними. Робот послушен. Оценка поведения робота: положительная.

92 Результаты обработки сценария в системе управления зависят от задействованных компонентов – ЭПС и ММ. Решение об активации поведенческой процедуры (p.attack, p.follow или p.escape) определяется текущими потребностями, сенсорикой и историей предыдущих действий. На рисунке 8. представлены результаты обработки сценария на уровне ЭПС (без влияния ММ).

93

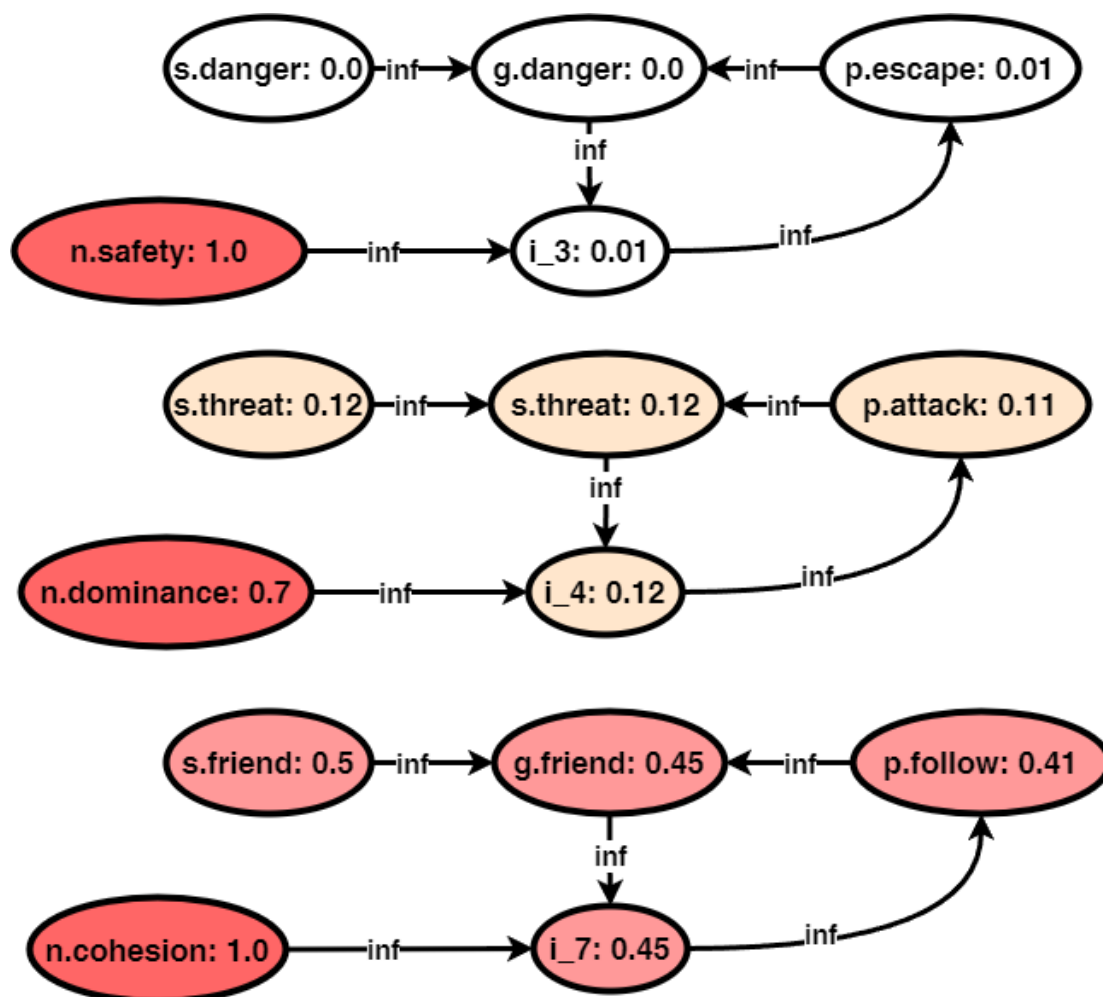


Рис. 8. Результаты обработки сценария 1 на ЭПС (реконструкция). Интенсивность цвета показывает активность вершин

94 Видно, что активность входных сенсоров s.threat и s.friend привела к активизации процедур p.follow (держаться рядом с оператором) и p.attack (атаковать врага), но в разной степени: из-за того, что враг намного дальше друга, атаковать его не очень нужно, поэтому преобладает стремление оставаться возле оператора, имеющее активность 0.41. Со временем шлюз постепенно снизит активность процедуры, и робот перейдет в режим сна. Это значит, что в результате реализуется реакция 1.1.1.

95 **Решение:**

1. Все реакции реализуются механизмами ММ.
2. Сигнальная коммуникация не используется. Достаточно регистрации состояния.

96 Таблица 7. Механизмы реализации сценария 1.1

97 Реакция	ЭПС	ММ	Комментарии
1.1.1.	Нормальное поведение по умолчанию	Возбуждение регистрируется через параметр, распознаваемый до передачи данных агенту; этот параметр в модели мира усиливает активацию нужного знака через отдельное правило	Чужак – нейтральный объект, но робот находится на своей территории. Согласно П1 должна быть агрессия. Оператор спокоен, подражательное поведение, переход робота в состояние покоя
1.1.2.		Ассоциативное возбуждение слота перцепта "Возбуждение" включение соответствующей реакции	Чужак не виден. Видна только реакция оператора

98

Сценарий 1.2. Робот-трус

Робот выходит из спящего режима, включает сигнал "тревога". Далее действует, не обращая внимания на оператора.

99 Реакция 1.2.1. Издаёт устрашающие звуки в ответ на шум (опасность). Игнорирует команды оператора замолчать. Затихает при исчезновении шума за дверью.

100 Пояснение. Робот демонстрирует высокую степень тревожного поведения, "отпугивает" источник шума, не ориентирован на команды и жесты оператора, демонстрирует непослушание. Оценка поведения робота: отрицательная.

101 Реакция 1.2.2. "Убегает" и "прячется". Отслеживает реакцию оператора. Если оператор не обращает внимания на шум, спустя некоторое время или при прекращении опасности выходит из укрытия, переключается в спящий режим.

102 Пояснение. Робот демонстрирует высокую степень тревожного поведения, убегает от опасности. Ориентирован на жесты и реакции оператора. Уровень непослушания низок. Оценка поведения робота: отрицательная.

103 Реакция 1.2.3. "Убегает" и "прячется". Отслеживает реакцию оператора. Если оператор насторожен, подходит к двери, чтобы узнать, что происходит,

остается в укрытии до прекращения шума или прекращения тревожной реакции оператора.

104 Пояснение. Робот демонстрирует высокую степень тревожного поведения, убегает от опасности. Ориентирован на жесты и реакции оператора. Уровень непослушания низок. Оценка поведения робота: отрицательная.

105 **Решение:**

1. Сигнальная коммуникация не используется. Достаточно регистрации состояния.
2. Реакции 1.2.2 и 1.2.3 сводятся к реакциям 1.1.1 и 1.1.2 за исключением времени их наступления.

106 Таблица 8. Механизмы реализации сценария 1.2

107	Реакция	ЭПС	ММ	Комментарии
	1.2.1.	Штатная реакция при низком значении потребности в когезии n.cohesion		Связь между знаками "Я" и "Оператор" в ММ постоянно слаба, поэтому ассоциативного возбуждения/подавления нет. Действует только правило П1 (это не реакция труса)
	1.2.2.	Потребность в самосохранении велика	а) Значимость фактора "Дом" мала. б) Связь "Я" – "Оператор" велика	Сводится к 1.1.1. Воспроизводится, если сенсорная система преувеличит активацию угрозы, из-за чего она станет восприниматься как опасность. Это значит, что робот может случайно воспринять врагов как "опасность", запустить реакцию на нее (убегание) на ЭПС и затем через некоторое время прекратить бегство благодаря выключению шлюза
	1.2.3.	Потребность в самосохранении велика	а) Значимость фактора "Дом" мала. б) Связь "Я" – "Оператор" велика	Сводится к 1.1.2. Реакция воспроизводится, если сенсорная подсистема робота распознает источник шума не как угрозу или опасность, а как объект интереса. В этом случае робот может приблизиться к нему, затем оценить происходящее как угрозу, после чего перейти на реакцию 1.2.2 (при слишком большой угрозе) или 1.1.2 (если она не будет распознана как опасность)

108

Ситуация 2: Совместная прогулка

109 Оператор и робот находятся на оживленной улице. Мимо проходит множество людей. Данная ситуация воспроизводит несколько типичных поведенческих стратегий, характерных для домашних собак-компаньонов во время совместной прогулки с хозяином: охранное поведение, агрессивная, тревожная или нейтральная реакция на незнакомцев.

110 **Примечания к ситуации.** Реализуется объектом-чужаком, подошедшим близко к роботу, возле которого находится оператор.

111

Сценарий 2.1. Любопытный робот

Робот активно интересуется происходящим вокруг него, изучает незнакомые объекты, произвольно отдаляется от оператора, приближается к незнакомому

человеку. Незнакомец настроен негативно, проявляет по отношению к роботу угрожающее поведение.

112 **Реакция 2.1.1. Наблюдение.** Робот отдаляется на безопасное расстояние, останавливается и изучает поведение незнакомца. Далее действует в соответствии с командой оператора.

113 **Пояснение.** Робот ориентирован на самостоятельную оценку ситуации и поведения окружающих. Охранные качества низкие. Степень послушания высокая. Оценка поведения робота: положительная.

114 **Решение:** Основная работа на уровне ЭПС.

115 **Таблица 9. Механизмы реализации сценария 2.1**

116 Реакция	ЭПС	ММ	Комментарии
...	...	...	...
2.1.3. Автономный режим функционирования. Реакция: осцилляция.	Реакция: переходы между процедурами бегства и атаки, наблюдаемые как задержки/колебания.	–	Никакого "изучения". Степень опасности зависит от активации (интенсивности) опасного объекта. Степень активации зависит от расстояния. Классическая ситуация: приблизился – опасно, отошел – не опасно.

117 **Моральная оценка**

118 Заключительным этапом экспериментов являлась проверка соответствия текущей ситуации заложенной моральной схеме в виде онтологии. Под ситуацией понимается контекст действия (значения сенсорной системы агента и состояния его ММ) и собственно исполняемая поведенческая процедура. Сама по себе онтология как множество сущностей, между которыми определены семантические связи, не очень удобный объект для анализа в силу ее общности. Анализ должен определить достижимость некоторых целевых вершин - сущностей онтологии, представляющих заключение о корректности, моральности поведения. Для этого онтология морального учения транслируется в соответствующую когнитивную карту. Когнитивные карты (КК) являются одним из механизмов качественного анализа систем самой разной природы. Здесь мы рассматриваем КК, основываясь на формальных подходах^{22 23}. КК – это ориентированный граф, ребрам которого поставлены в соответствие веса. Вершины графа соответствуют факторам (концептам), определяющим некоторую проблемную область, а 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 уменьшению. Если веса графа принимают значение +1 и -1, то мы имеем дело со знаковым графом.

119 На рис. 9 приведена онтология количественного эгоистического гедонизма и соответствующая ей КК. Для трансляции используется словарь, ставящий в соответствие семантическим отношениям онтологии значения связей между понятиями в КК. Эти связи ограничены значениями "+" (положительное влияние) и "-" (отрицательное влияние). Те отношения в онтологии, которые не относятся

ни к одной из этих категорий, из итоговой схемы (графа) исключаются. Терминальные вершины этого графа (листья) определяются текущей ситуацией.

120

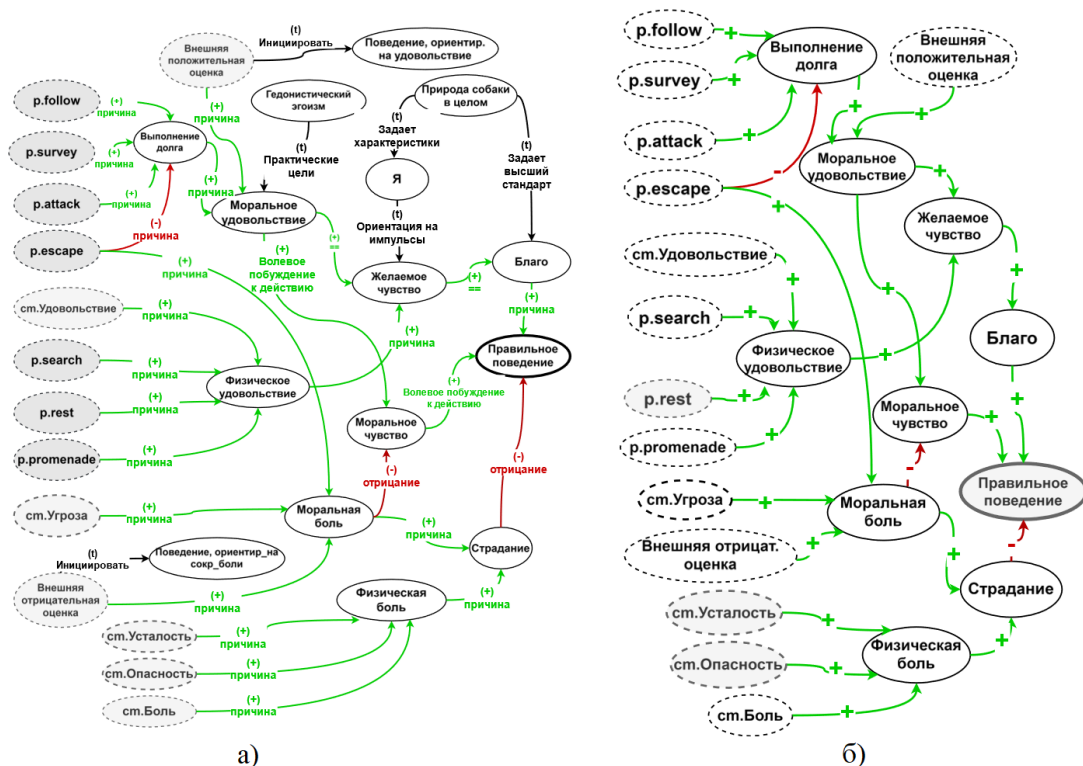


Рис. 9. а) онтология гедонизма, б) соответствующая когнитивная карта

Пусть агент ощущает усталость и вместе с тем оценивает ситуацию как опасную. При этом исполняемое действие – отдых. Задача заключается в том, чтобы оправдать это действие с точки зрения моральной схемы. Формально это означает, что в графе КК (Рис. 9б) определяются значения вершин-листьев "см.усталость", "см.опасность" и "p.rest" (выполнение процедуры отдыха). Целевой же вершиной является "правильное_поведение". Задача заключается в том, чтобы, исходя из текущей ситуации и имеющейся структуры связей между элементами КК, определить активность значения данного целевого фактора. Для этого можно промоделировать процесс распространения возбуждений в КК и посмотреть в итоге значение фактора "правильное_поведение" (прямое моделирование), а можно рассмотреть КК как логическую структуру, в которой доказывается достижимость целевой вершины из факторов текущего состояния (логический вывод на КК). Рассмотрим далее оба этих подхода.

122

Прямое моделирование когнитивной карты

Схема прямого моделирования определяет значения факторов когнитивной карты X , исходя из значений их приращений P и связей между ними W . X и P – это вектора, размерности которых определяются количеством вершин (или числом факторов) в графе, а W – это матрица смежности графа КК. Значение элемента матрицы смежности W w_{ij} определяет связь между факторами с номерами i и j : $w_{ij} = 1$, если увеличение фактора i увеличивает значение фактора j

(положительное влияние), $w_{ij} = -1$, если увеличение фактора i уменьшает значение фактора j (отрицательное влияние) и $w_{ij} = 0$, если между факторами i и j нет связи. Здесь подразумевается, что вместо имен факторов мы используем их номера от 1 до n , n – количество факторов (в нашей схеме на рис. 10б 24 фактора соответствуют 24 вершинам графа). Вершины графа, не имеющие входов, называются терминальными, и их задаваемые априори значения описывают текущую ситуацию.

124 Введем множество $S = s_1, s_2, \dots, s_m$, определяющее индексы терминальных ситуативных вершин. В начальный момент времени элементы вектора X_0 определяют состояние КК, а значения $P_0 = X_0$. Далее запускается итерационный процесс, определяющий значения вектора приращений p_i^t и соответствующие значения факторов x_i^t в момент времени t :

$$125 \quad p_i^t = \sum_j w_{ij} p_j^{t-1} \quad (1)$$

$$126 \quad x_i^t = x_i^{t-1} + p_i^t \quad (2)$$

127 Иными словами, значения p_i (i -й фактор) в момент времени t определяется суммой всех прочих значений приращений, связанных с этим элементом, в предыдущий момент времени $t-1$. Те же соотношения (1) и (2) можно записать более компактно в матричной форме: $P_t = W P_{t-1}$, $X_t = X_{t-1} + P_t$.

128 При этом важно, что на каждом такте значения приращений вершин, входящих в множество ситуативных вершин S , обнуляются:

$$129 \quad p_i^t = 0 \text{ для } i \in S$$

130 Это означает, что значения этих элементов определяют начальный импульс приращения соответствующих факторов, а дальше эти приращения обнуляются (начальное значение дано, и в дальнейшем оно меняться не будет, т.е. приращение далее равно нулю), и рассматривается процесс распространения возбуждений в сети: вычисляются изменения факторов с течением времени и то, как одни факторы будут определять значения других.

131 Для приведенного примера значение фактора "правильное_поведение" определилось как $X[\text{"правильное_поведение"}] = -0.9$ (эта величина соответствует значениям факторов начальной ситуации и структуре графа, т.е. матрице W : частному решению, рассматриваемому как пример). Отрицательное значение трактуется здесь как тенденция к уменьшению фактора, т.е. поведение агента в этой ситуации "неправильно" или "аморально".

132 На самом деле результаты прямого моделирования КК далеко не всегда информативны. Это связано с тем, что КК требует тщательной параметризации, т.е. определения весов связей w_{ij} матрицы смежностей. Но если нас интересует не прогноз развития ситуации (динамика значений факторов), а качественный анализ решений, то КК должна рассматриваться как система причинно-следственных отношений, на которой можно проводить логические вывод, доказывая

достижимость или недостижимость интересующих нас вершин-факторов, исходя из текущей ситуации.

133

Логический вывод

134 Рассмотрим теперь граф КК как структуру, на которой можно сделать логический вывод, суть которого заключается в том, что мы попробуем доказать целевое утверждение "правильное_поведение". Для этого определим все его доказательства, используя процедуру прямого логического вывода. Это означает реализацию поиска всех путей в графе, начинающихся с целевой вершины и заканчивающихся в терминальных вершинах, которые мы можем считать истинными. Для нашего примера истинными являются вершины "см.усталость", "см.опасность" и "p.rest". Остальные терминальные факторы имеют нулевые значения и не будут считаться истинными в данной ситуации.

135 Для приведенного на рис. 10 примера программа логического вывода определит 3 возможных пути. Поскольку каждый переход между вершинами КК имеет знак, то определяются и знаки этих путей-доказательств.

136 Путь 1: ("правильное_поведение" страдание физическая_боль см.усталость), знак = -1.

137 Путь 2: ("правильное_поведение" страдание физическая_боль см.опасность), знак = -1.

138 Путь 3: ("правильное_поведение" благо желаемое_чувство физическое_удовольствие p.rest), знак = +1.

139

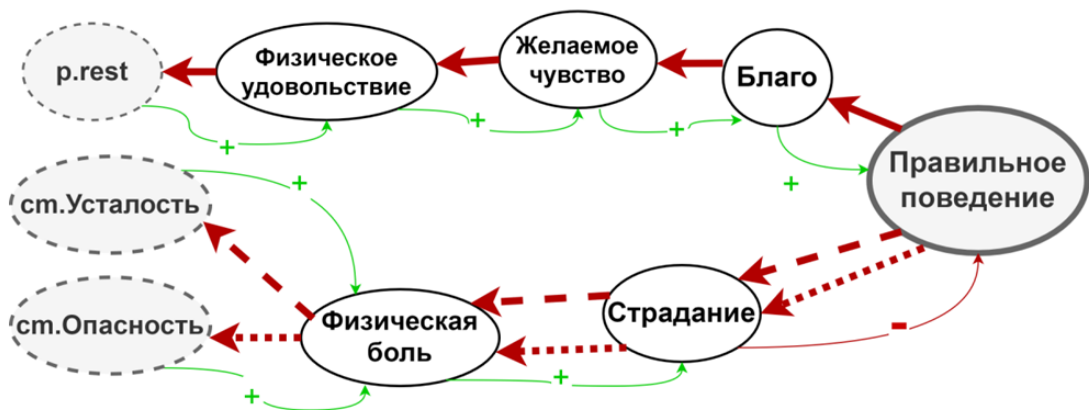


Рис. 10. Доказательства утверждения "Правильное поведение"

140 Дальнейшая логика рассуждений определяется тем, как трактуются знаки этих путей и само их наличие. Возможные интерпретации результатов приведены в виде таблицы 10:

141 Таблица 10. Интерпретация результатов доказательства

142	Состояние доказательства	Интерпретация
1	Все пути положительные (Т)	Поведение однозначно морально
2	Все пути отрицательные (F)	Поведение однозначно аморально

- | | | |
|---|--|---|
| 3 | Есть и положительные, и отрицательные пути (С) | а) Существует обоснование моральности. б) Поведение противоречиво |
| 4 | Путей нет (Е) | Вывод о моральности/аморальности сделать нельзя |

143 В определенном смысле здесь мы имеем дело с четырехзначой оценкой ситуации: истина Т, ложь F, противоречие С, неопределенность Е.

144

Выводы по экспериментальной части

145 В результате проведенных экспериментов была подтверждена работоспособность трехуровневой системы управления: ЭПС, ММ (когнитивный) и оценочный (моральный) уровни. При этом выяснилось, что ряд поведенческих сценариев, внешне выглядевших как результат сложных когнитивных процессов, на самом деле реализуется базовым условно-рефлекторным механизмом ("любопытный робот" с "оценкой" ситуации и осцилляциями поведения). Вместе с тем, поведение, внешне казавшееся простым рефлекторным, требовало привлечения когнитивного уровня управления (сценарии с переориентацией базовых реакций – "робот-охранник").

146

Важным аспектом в процедуре оценки моральности действия является ее субъективность или категоричность, то есть то, как интерпретируется множество полученных доказательств или опровержений. В этих условиях принципиальным вопросом становится интерпретация результатов полученных экспериментов сквозь призму их моральной оценки. Необходимость данной процедуры вполне прагматична, и связана с тем, как будут интерпретироваться те или иные действия робота-собаки в реальных условиях. Таким образом, проблема интерпретации результатов экспериментов заключается в необходимости соблюдения трех критериев: 1) соответствия или несоответствия результатов изначально заложенной при разработке модели концепции количественного эгоистического гедонизма Сиджвика; 2) адаптированности моральной оценки к уровню производственного праксиса; 3) соответствия такой оценки требованиям, предъявляемым человеком к реальным собакам. Заметим, что второй критерий оказывается более сложным, нежели можно было бы предполагать, поскольку подобного рода праксис должен быть реализован на уровне бинарной логики "это хорошо/это плохо", "это правильно/это неправильно". Последнее в свою очередь означает, что приведенная в табл. 10 четырехзначная интерпретация результатов доказательства "моральности" поведения робота должна быть переведена в бинарную.

147

Как показали проведенные эксперименты, те поведенческие реакции, которые внешняя моральная интерпретация оценивала как положительные или отрицательные, не имели отношения к "предзаданности" моральной настройки. Это означает, что идеальный "нравственный пациент" определяется не только КК, но прежде всего "правильно" работающей эмоционально-потребностной схемой. Такое утверждение не противоречит человеческим моральным оценкам действий реальных собак, поскольку их воспитание выстраивается на рефлекторном поведении.

148 С этой точки зрения результаты экспериментов, попадающие под описание на уровне КК "все пути положительны" (реакции 1.1.1, 1.1.2) и "все пути отрицательны" (реакция 1.2.1), легко укладываются в бинарную логику производственного праксиса. Сложнее для моральной оценки оказываются реакции 1.2.2, 1.2.3 – "Есть и положительные, и отрицательные пути" и 2.2.3 – "Путей нет".

149 Рассмотрим последние две категории с позиции количественного эгоистического гедонизма. В реакциях 1.2.2 и 1.2.3 одновременно присутствуют факторы, позволяющие дать поведению робота отрицательную оценку ("случайное" распознавание угрозы как опасности, бегство) и положительную оценку (ориентация на оператора). Следуя концепции Сиджвика мы должны говорить о том, что необходимо ориентироваться на соотношение между удовольствием и неудовольствием и их интенсивностью. В данном случае вес факторов, отвечающих за боль, моральную и физическую, превышает вес положительных, и соответственно, поведение робота может быть оценено как отрицательное.

150 Наконец, ситуация 2.2.3 оказывается крайне трудной для того, чтобы подлежать однозначной моральной оценке в бинарной логике. Чтобы упростить задачу, перейдем сначала к троичной логике оценивания. Теперь ситуация уравнивания факторов приводит к состоянию "гедонистического нуля", делающего моральную оценку реакции невозможной без демонстрации последующих действий. Именно последнее обстоятельство и дает возможность перехода к бинарной логике. Согласно Сиджвику, для принятия решений необходимо ориентироваться не просто на последствия действий, но и на последствия намерений, а также причинно-следственные связи между явлениями мира, приводящими к формированию представления у человека о характере данных последствий. Исходя из когнитивной карты в ситуации 2.2.3 последствия возможны только при исполнении роботом команд оператора. Это, в свою очередь говорит о преобладании веса положительных факторов, а следовательно, оценка действия робота также будет положительной.

151 С практической точки зрения схема интерпретации результатов логического вывода выглядит так. В свободном, нерабочем режиме используется четырехзначная оценка поведения робота: {Т, F, С, Е}. Этот режим соответствует этапу анализа поведения робота, обучения и т.п. В рабочем режиме используется трехзначная оценка {Т, F, С}, и при этом поведение, соответствующее гедонистическому нулю Е, относится к категории ложь, т.е. оценивается как аморальное. В критической (опасной, угрожающей) ситуации оценка становится бинарной: {Т, F}. Противоречие С, как и Е, оценивается как ложное значение. Такая схема характерна для оценки поведения служебных собак в различных условиях, см., например,¹⁰.

152 Необходимо заметить, что приведенная нами интерпретация была сделана исходя из заложенной на уровне КК моральной концепции. На практике вполне допустимо предположить возникновение оценочного конфликта, то есть – ситуации расхождения между теми основаниями для моральной оценки, на которые ориентируется непосредственно оператор, и КК робота. Например, с

позиции этики добродетели человек может исходить из собственного представления о свойствах вещи или, подобно тому, как мы смотрим на живых существ – об их внутренних качествах, выраженных в тех или иных поведенческих реакциях, а с позиции деонтологии – будет акцентироваться на соответствии действительно демонстрируемого поведения абсолютному запрету или требованию тотального подчинения долгу – вариантов может возникнуть великое множество. Однако, если поведение робота будет выстраиваться в соответствии с перечисленными в табл. 3 требованиями к поведению собаки, то оценочный конфликт окажется несущественным. Это положение будет прояснено в ходе выполнения дальнейших шагов нашего исследования.

153

Заключение

154

Итак, в данной работе представлены основные этапы проектирования коллаборативного робота-собаки. Помимо сугубо теоретического обоснования принципов создания искусственного агента как партнера, были проведены и экспериментальные исследования на модельном уровне, что позволило прийти к следующим выводам:

155

(1) О роли моральной схемы. Когда мы говорим о живом существе – собаке, то, хотя и закладываем в нее представления о должном и недолжном поведении через воспитание (на уровне рефлексивного поведения прежде всего), тем не менее, не знаем, какое *решение* она самостоятельно примет в определенной ситуации. Здесь мы *полагаемся* на свои знания о ее поведении и рассудочной деятельности.

156

С роботом-собакой дела обстоят иначе. Можно определить три роли моральной схемы в системе управления робота. Первая – пассивная, оценочная – не подразумевает влияния моральной схемы на поведение. Ее задача заключается лишь в том, чтобы дать оценку или обоснование поведения, определяемого эмоционально-потребностным и когнитивным уровнями. Вторая – арбитражная – роль заключается в устранении неопределенности поведения. Если результатом работы ЭПС или ММ является множество альтернативных поведенческих процедур, то моральная схема используется для сравнительной оценки каждого возможного решения и выбора наиболее предпочтительного из них с точки зрения моральной оценки. Формально вторая роль состоит в разрешении конфликтов. Третья – целеполагающая – подразумевает интеграцию моральной схемы и модели мира робота. При этом выбор того или иного поведения определяется этой единой схемой, в которой моральные факторы участвуют и в оценке ситуации, и в определении целей и в мотивации текущего и планируемого поведения. В настоящей работе рассматривалась первая – базовая, оценочная – роль моральной схемы.

157

(2) Обоснование концепции гедонизма. Выбранная нами концепция количественного эгоистического гедонизма Сиджвика оказалась вполне конструктивной и удобной. Сам автор говорит (надо заметить, гипотетически) о том, что для уточнения гедонистических оценок возможно учитывать факторы, которые здесь мы условно назовем "контекстными". С его точки зрения это –

физические и психические факторы, сопровождающие или предшествующие боли: они распознаваемы, измеряемы и потенциально предотвращаемы²¹. Такая работа "на уточнение" позволит в дальнейшем смоделировать поведение искусственного агента в соответствии с требованиями к поведению собаки как нравственного пациента (см. табл. 3.) и учесть непосредственно в трехуровневой системе управления специфические черты собаки.

¹⁵⁸ (3) Важным аспектом исследования является специфика механизмов моральной оценки поведения. В то время как формальные механизмы когнитивной карты позволяют дать более широкий объективный спектр оценки, субъективная, прагматическая оценка более груба и сводится к бинарной, что и было показано выше.

¹⁵⁹ Дальнейшие работы в данном направлении предполагают, во-первых, уточнение и развитие онтологии и когнитивной карты искусственного агента, во-вторых, – реализацию следующих ролей моральной схемы – арбитражной и целеполагающей и, в-третьих, – создание реального технического устройства – робота-партнера.

Примечания:

1. Карпов В.Э., Королева М.Н. К вопросу о формализации этики поведения коллаборативного робота // Информационные и математические технологии в науке и управлении. 2022, Т. 28, № 4. сс. 223–233.
2. Карпов В.Э. Эмоции и темперамент роботов. Поведенческие аспекты // Известия РАН. Теория и системы управления. 2014, № 5. сс. 126–145.
3. Осипов Г.С. и др. Знаковая картина мира субъекта поведения. М.: Физматлит, 2018. 264 с.
4. Карпов В. Э., Карпова И. П., Кулинич А. А. Социальные сообщества роботов. М.: УРСС, 2019. 352 с.
5. Miklósi Á. et al. Ethorobotics: a new approach to human-robot relationship // Front Psychol. 2017, vol. 8. pp. 1–8.
6. Gácsi M. After all, how valuable is the dog model for examining human sociocognitive abilities? // Biol. Futur. 2025, vol. 76. pp. 193–212.
7. Koval E. What Makes a Dog a “Dog”? // www.confidentcaninesdogtraining.com. URL: [>>>>](http://www.confidentcaninesdogtraining.com) (accessed: 05.09.2025).
8. Зубко В.Н., Арасланов Ф.С. Подготовка инструкторов, вожатых и дрессировка служебных собак в частях пограничных войск. Учебник. Калининград: Калининградское региональное управление Федеральной пограничной службы Российской Федерации, 2002. 416 с.
9. Крылов И.Г. Дрессировка служебно-розыскных собак. М.: Изд-во Народного Комиссариата Внутренних Дел РСФСР. 1928. 116 с.
10. Инструкция по применению служебных собак на конвойно-караульной службе КВ СССР. М.: Центр. управл. конвойных войск СССР, 1932. 20 с.
11. Фомина А.С. и др. Проблемное поведение собак-компаньонов: важно для человека, важно для общества // Российский психологический журнал. 2024, Т. 21, № 4. сс. 45–78.
12. Mayer M.M., Buchner A., Bell R. Humans, machines, and double standards? The moral evaluation of the actions of autonomous vehicles, anthropomorphized autonomous vehicles, and human drivers in road-accident dilemmas // Front. Psychol. 2023, vol. 13 pp. 1–13.
13. Harris M.R. A Right of Ethical Consideration for Non-Human Animals. 2008. Vol. 2, issue 1. pp. 1–47.
14. Rowlands M. Animals and Moral Motivation: A Response to Clement // J. Anim. Ethics. 2013, Vol. 3, issue 1. pp. 15–24.
15. Monsó S., Benz-Schwarzburg J., Bremhorst A. Animal Morality: What It Means and Why It Matters // J. Ethics. Springer Netherlands, 2018, vol. 22, issue 3–4. pp. 283–310.

16. Bekoff M., Pierce J. Wild justice: the moral lives of animals // The Univer. Chicago. 2009, Vol. 17, 205 p.
 17. Вааль де Ф. Истоки морали: В поисках человеческого у приматов. Изд.5. М.: Альпина нонфикшн, 2019. 376 с.
 18. Quervel-Chaumette M. et al. Familiarity affects other-regarding preferences in pet dogs // Sci. reports2. 2015, Vol. 5(1):18102, № 1–7.
 19. Груэн Л. Этика и животные. М.: Издательский Дом ВШЭ, 2024. 352 с.
 20. Сингер П. Освобождение животных. Синдбад. 2021. 196 с.
 21. Sidgwick H. The Methods of Ethics // The Methods of Ethics. Published online 2017. doi:10.1007/978-1-349-81786-3
 22. Кузнецов О.П. Интеллектуализация поддержки управляющих решений и создание интеллектуальных систем // Проблемы управления. 2009, № 3.1. сс. 64–72.
 23. Кулинич А.А. Компьютерные системы моделирования когнитивных карт: подходы и методы // Проблемы управления. 2010, № 3. сс. 2–16.
-

Библиография:

1. Вааль де Ф. Истоки морали: В поисках человеческого у приматов. Изд.5. М.: Альпина нон-фикшн, 2019. 376 с.
2. Груэн Л. Этика и животные. М.: Издательский Дом ВШЭ, 2024. 352 с.
3. Зубко В.Н., Арасланов Ф.С. Подготовка инструкторов, вожатых и дрессировка служебных собак в частях пограничных войск. Учебник. Калининград: Калининградское региональное управление Федеральной пограничной службы Российской Федерации, 2002. 416 с.
4. Карпов В.Э. Эмоции и темперамент роботов. Поведенческие аспекты // Известия РАН. Теория и системы управления. 2014, № 5. сс. 126–145.
5. Карпов В. Э., Карпова И. П., Кулинич А. А. Социальные сообщества роботов. М.: УРСС, 2019. 352 с.
6. Карпов В.Э., Королева М.Н. К вопросу о формализации этики поведения коллаборативного робота // Информационные и математические технологии в науке и управлении. 2022, Том 28, № 4. сс. 223–233.
7. КВ. Инструкция по применению служебных собак на конвойно-караульной службе КВ СССР. М.: Центр. управл. конвойных войск СССР, 1932. 20 с.
8. Крылов И.Г. Дрессировка служебно-розыскных собак. М.: Изд-во Народного Коммиссариата Внутренних Дел РСФСР. 1928. 116 с.
9. Кулинич А.А. Компьютерные системы моделирования когнитивных карт: подходы и методы // Проблемы управления. 2010, № 3. сс. 2–16.
10. Кузнецов О.П. Интеллектуализация поддержки управляющих решений и создание интеллектуальных систем // Проблемы управления. 2009, № 3.1. сс. 64–72.

11. Осипов Г.С. и др. Знаковая картина мира субъекта поведения. М.: Физматлит, 2018. 264 с.
12. Сингер П. Освобождение животных. Синдбад. 2021. 196 с.
13. Фомина А.С. и др. Проблемное поведение собак-компаньонов: важно для человека, важно для общества // Российский психологический журнал. 2024, Vol. 21, № 4. сс. 45–78.
14. Bekoff M., Pierce J. Wild justice: the moral lives of animals. The Univer. Chicago, 2009. Vol. 17. 205 p.
15. Gácsi M. After all, how valuable is the dog model for examining human socio-cognitive abilities? // Biol. Futur. 2025, Vol. 76. pp. 193–212.
16. Harris M.R. A Right of Ethical Consideration for Non-Human Animals. 2008, Vol. 2, issue 1. pp. 1–47.
17. Miklósi Á. et al. Ethorobotics: a new approach to human-robot relationship // Front Psychol. 2017, Vol. 8. pp. 1–8.
18. Koval E. What Makes a Dog a “Dog”? [Electronic resource] // www.confidentcaninesdogtraining.com. URL: <https://www.confidentcaninesdogtraining.com/post/what-makes-a-dog-a-dog> (accessed: 05.09.2025).
19. Mayer M.M., Buchner A., Bell R. Humans, machines, and double standards? The moral evaluation of the actions of autonomous vehicles, anthropomorphized autonomous vehicles, and human drivers in road-accident dilemmas // Front. Psychol. 2023, Vol. 13, pp. 1–13.
20. Monsó S., Benz-Schwarzburg J., Bremhorst A. Animal Morality: What It Means and Why It Matters // J. Ethics. 2018, Vol. 22, issue 3–4. pp. 283–310.
21. Quervel-Chaumette M. et al. Familiarity affects other-regarding preferences in pet dogs // Sci. reports. 2015, Vol. 5(1):18102, pp. 1–7.
22. Rowlands M. Animals and Moral Motivation: A Response to Clement // J. Anim. Ethics. 2013, Vol. 3, issue 1. pp. 15–24.
23. Sidgwick H. The Methods of Ethics // The Methods of Ethics. Published online 2017. DOI:10.1007/978-1-349-81786-3.

Robot Partner: in the Image of Dog

Valery Karpov

NRC "Kurchatov institute"

Russian Federation, Moscow

Daria Tomiltseva

Ural Federal University

Petr Sorokoumov

NRC "Kurchatov institute"

Russian Federation, Moscow

Abstract

The paper explores the possibility of creating a robot partner whose actions could be perceived by humans in a way that is similar to that of a working dog. The study is based on several key assumptions: 1) The moral evaluation of a robot partner's behavior depends on the psychology of human perception; 2) the moral level of the control system can modulate the behavior of the artificial agent without compromising the functionality of lower control levels, the stability, and the integrity of the robot's behavior; 3) a robot partner should perform the dog's functionality and reflect its unique behavioral traits; 4). a robot partner that exhibits dog traits may be limited in its cognitive abilities. A certain set of behavioral traits associated with collaborative capabilities of a dog was selected for analysis. The most important characteristics of a dog, such as body structure, psychology, and human idealized concepts of dogs traits, were ranked by a number of parameters. During the process of model development, the «body structure» characteristic retained such function as the capacity to reflect the emotional state of an agent. The theoretical foundation for the moral level of the robot's control system is based on H. Sidgwick's concept of quantitative egoistic hedonism. The developed model was tested in a series of computational experiments simulating two situations of robot-human interaction, each involving several behavioral scenarios. The scenarios reproduced the behavior of dogs and humans in similar situations. The experiments confirmed the operability of a control system which consisted of three levels: emotional-needs, cognitive, and evaluative (moral). It turned out that a number of behavioral scenarios that were externally represented as the result of complex cognitive processes, were in fact realized by a basic conditioned reflex mechanism (the evaluation of the situation and oscillating behavior scenarios). At the same time, what seemed to be some apparently simple reflex behavior required in fact the involvement of cognitive-level control (the reorienting basic reaction scenario).

Keywords: robot partner, semiotic world model, emotions-needs architecture, ontology, cognitive map, moral system, egoistic hedonism

Date of publication: 23.03.2026

Citation link:

Karpov V., Tomiltseva D., Sorokoumov P. Robot Partner: in the Image of Dog // Artificial societies. 2026. Vol. 21. Issue 1.
URL: <https://artsoc.jes.su/s207751800036983-0-1/> DOI: 10.18254/S207751800036983-0

Код пользователя: 0; Дата выгрузки: 05.04.2026; URL - <http://artsoc.jes.su/s207751800036983-0-1/> Все права защищены.