Российская ассоциация искусственного интеллекта
Санкт-Петербургский институт информатики и автоматизации
Российской академии наук

ПЯТЫЙ ВСЕРОССИЙСКИЙ НАУЧНО-ПРАКТИЧЕСКИЙ СЕМИНАР «БЕСПИЛОТНЫЕ ТРАНСПОРТНЫЕ СРЕДСТВА С ЭЛЕМЕНТАМИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА» (БТС-ИИ-2019)

22-24 мая 2019
г. Санкт-Петербург, Россия

Труды семинара

УДК 004.8, 17, 519.7

ОТ ПОДРАЖАТЕЛЬНОГО ПОВЕДЕНИЯ К ЭМПАТИИ В СОЦИУМЕ РОБОТОВ

В.Э. Карпов (karpov-ve@yandex.ru)
Национальный исследовательский центр "Курчатовский институт", Москва
МФТИ

Аннотация. В работе рассматриваются некоторые конструктивные аспекты явного введения в систему управления роботом понятия субъективного Я. Показано, что эта компонента позволяет естественным образом реализовать такие феномены, как подражательное поведение, социальное обучение и эмпатия. В основе разработанных моделей лежит понятие степени близости наблюдаемого контрагента к субъекту. Делается предположение, что эти модели могут стать основой для создания такого механизма адаптивного поведения, который называется моралью.[1]

Ключевые слова: групповая робототехника; социальные сообщества роботов; подражательное поведение; эмпатия; мораль; субъективное Я.

Введение

Одним из направлений групповой робототехники являются т.н. модели социального поведения. Ключевой задачей этого направления является создание конструктивных моделей, позволяющих реализовывать феномены социальной организации в группах искусственных агентов [Карпов, Карпова, Кулинич, 2019]. Основная идея парадигмы моделей социального поведения (МСП) заключается в том, чтобы рассматривать принципы организации сообществ роботов с точки зрения некоторого универсального адаптационного механизма. Сам вопрос о видах поведения, в т.ч. – социального – является до сих пор открытым. Так, с одной стороны психологии и биологи говорят о четырех основных формах поведения: пищевое, оборонительное, половое и исследовательское, см., например, [Лурия, 2007]. С другой стороны, те же этологи или специалисты в области социальной биологии выделяют дополнительные, специфические формы и

[1] Работа выполнена при частичной поддержке грантов РФФИ 17-29-07083 офи_м (семиотические аспекты) и 16-29-04412 офи_м (эмпатия, модели поведения).

модели поведения, необходимые для организации социума, такие, как контагиозное (заразное), подражательное, агрессивное и пр. [Зорина, Полетаева, Резникова, 2002]. Многие из этих механизмов достаточно сильно пересекаются друг с другом, зачастую представляя собой лишь некоторое условное обозначение проявления тех или иных базовых механизмов. Так, возникает сомнение в существовании агрессивного поведения как самостоятельной, специфичной сущности (см. [Карпова, Карпов, 2018]), хотя о самой агрессии написано множество фундаментальных трудов (правда, чаще представляющих объемное описания тех или иных феноменов и проявлений того, что мы называем агрессией, см. [Лоренц, 1994], [Ениколопов, Кузнецова, Чудова, 2014]) и др.

Важным вопросом парадигмы МСП является определение соответствия уровня развития архитектуры искусственного агента той или иной форме или механизму социального поведения. Такие механизмы, как когезия (стремление держаться вместе), обучение (в простейших формах), доминирование, дифференциация функций и некоторые другие вполне реализуемы на рефлекторном уровне организации системы управления. Напротив, механизмы социального обучения и подражательное поведение – это уже прерогатива когнитивной архитектуры. Здесь мы не будем касаться вопросов организации когнитивного уровня агентов. Эти вопросы относятся к исследованиям в области BDI-архитектур (см., например, [Кулинич, 2018]) и в определенном смысле – прикладной семиотики [Осипов и др., 2018]. Мы затронем в этой работе лишь один аспект когнитивных архитектур – элемент, называемый "субъективное Я". Далее будет показано, как введение этого элемента позволяет реализовать схему подражательного поведения, а также реализовать модель социального обучения. Кроме того, наличия этого "субъективного Я" позволяет подойти к решению вопросов эмпатии и даже некоторых аспектов этического поведения.

1. Субъективное Я

В социальной психологии субъективное Я (далее – С.Я.) – это то, что человек сам о себе думает и знает. С.Я. называют также Я-познающее. Более строго, в психологии С.Я. состоит из знаний трех типов, которые вытекают из активной деятельностной роли личности и включают такие компоненты, как: (1) континуальность, т.е. осознание себя одним и тем же человеком в течение онтогенеза; (2) знания об индивидуальности, о своем отличии от других; (3) воля, чувство личного контроля [Знаков, 2005]. В прикладной семиотике существует аналогичное понятие – субъект деятельности. Включение этого понятия в модель мира формирует т.н.

картину мира [Осипов и др., 2018]. При этом полагается, что одной из основных функций субъекта деятельности является целеполагание.

Следует отметить, что чаще всего, вне рамок семиотического подхода, С.Я. определяется неявным образом. В любом случае оценка близости текущего состояния к целевому, определение местоположения и пр. – все это происходит относительно самого агента. Явное же задание компоненты С.Я. не только позволяет унифицировать описание картины мира, но, прежде всего, дает возможность создавать модели различных форм поведения, описывающих взаимодействие между агентами.

Прежде, чем будут рассмотрены конструктивные аспекты применения С.Я. для описания некоторых форм поведения, сделаем важное замечание. Далеко не все модели социального поведения требуют введения компоненты С.Я., несмотря даже на внешнюю сложность и кажущуюся необходимость наличия семиотической модели. Таким примером является т.н. контагиозное ("заразное") поведение.

Контагиозное поведение. Суть этого поведения достаточно проста: в отсутствии явного стимула, регистрируемого сенсорной системой индивида, за счет внешнего сигнала могут быть запущены его различные поведенческие реакции. Например, сигнал опасности подхватывается остальными членами группы, заставляя стаю птиц сниматься с места, даже если птицы не видят непосредственной угрозы. Это – простейший тип сотрудничества, реализация принципа "делай, как я" [Тинберген, 1993]. Важно, что это не подражание, так как реагирующие особи не научаются выполнять определенные движения, наблюдая за действиями других. В [Карпов, Карпова, Кулинич, 2019] показано, как реализуется контагиозное поведение без привлечения С.Я., а также понятия "Иной" ("Ты").

Подражательное поведение. Реализация этого феномена требует ответа на следующие вопросы: (1) чему следует подражать, (2) почему это следует делать и (3) кому надо подражать. Будем называть агента, реализующего подражательное поведение, наблюдателем, а того, чьему поведению он подражает – контрагентом (в этологии – конспецификом).

Первый вопрос – это определение того, чему, собственно надо подражать. Т.е. необходимо понять, что делает контрагент или в каком состоянии он находится. Разумеется, определение или распознавание состояния агента – это крайне сложная задача, связанная с анализом сцен, причем зачастую – динамических. Однако здесь зачастую используется один чисто технический "трюк". Контрагент сам сигнализирует о том, в каком состоянии он находится в текущий момент времени. Именно так в некоторых экспериментах поступают роботы лаборатории робототехники НИЦ "Курчатовский институт". На роботах расположены ИК-маяки, выдающие в частотой 5-10 Гц десятиразрядные коды, определяющие идентификатор робота, его характеристики и код выполняемого в текущий

момент времени действия (состояние). Т.е. робот сообщает явно, что он "спит", "ищет пищу", "убегает" и т.п. Этот подход может иметь и некоторое биологическое основание. Итак, каждое действие имеет свое внешнее проявление.

Вопрос обусловленности подражания несколько сложнее. Будем считать, что элементы системы управления, образующие сеть, снабжены еще одним, помимо возбуждающего или инициирующего, дополнительным подтверждающим входом. Это – вход для сигнала от вершины С.Я. Так, действие не будет активировано, если не будет подтверждающего сигнала от С.Я., интерпретируемого как "принадлежность" этого действия агенту. В определенном смысле это – чувство (ощущение) самости, т.е. отождествление или восприятие объекта, как своего. Без такого ощущения в животном мире происходит рассогласование деятельности. Например, известно сложное психоневрологическое расстройство, называемое синдромом чужой руки. Одним из его клинических симптомов является наличие субъективных ощущений у пациента чужеродности конечности.

Ответ на вопрос "кому подражать" не является очевидным. Речь идет о том, что при наблюдении контрагента происходит не просто определение степени его похожести или близости к агенту. В простейшем случае так работает запаховая метка у муравьев, которые могут определить эту степень близости: от принадлежности контрагента к своему клану до определения его как совершенно чужого. Важнее то, что происходит отождествление наблюдаемого объекта (контрагента) с С.Я. Это означает, что при наблюдении "близкого" контрагента происходит то же подтверждение самости, что и при активизации вершины С.Я. Как происходит это определение степени "похожести" – не существенно. Похожесть чаще всего определяется набором наблюдаемых признаков. Итоговая схема модели подражательного поведения приведена на Рис. 6.

Пунктирная стрелка, ведущая к элементу "Сигнал", означает, что состояние вершины-источника может быть зарегистрировано внешним наблюдателем. Вершина "Наблюдение" отвечает за определение близости наблюдаемого контрагента. На этой схеме вершины s_i отвечают за наблюдаемый набор признаков, приводящих к активизации вершин-детекторов ("Опасность", "Угроза", "Пища") и далее – соответствующих процедур ("Бегство", "Нападение", "Транспортировка").

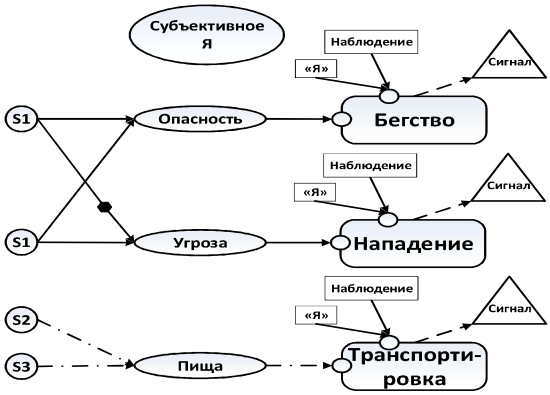

Рис. 6. Пример схемы подражательного поведения. Сплошные стрелки – это сформировавшиеся связи. Штрих-пунктирные стрелки – это формирующиеся в ходе подражания связи

Предположим, что у агента нет связей между наблюдаемыми признаками s_2, s_3 и выполнением действия "Транспортировка". Если агент видит s_2, s_3, а также то, что контрагент совершает действие "Транспортировка" (активен вход "Наблюдение"), то мы получаем систему с активными элементами, между которыми формируется ассоциативная связь. Агент приобретает новый навык.

Подчеркнем еще раз, что основное отличие контагиозного поведения от подражательного заключается именно в приобретении новых навыков, причем это становится возможным в силу того, что в схеме управления определяется вершина С.Я.

2. Социальное обучение

Компоненты С.Я. и "Наблюдение" позволяют описать такой интересный и важный феномен поведения, как социальное обучение (или обучение на чужих примерах). Социальное обучение – это термин, который используется в этологии по отношению к способностям животных приобретать опыт, связанный с взаимодействием с другими особями [Резникова, 2004]. Среди множества видов и форм проявления социального обучения в настоящей работе нас интересует то, которое связано с формированием условных рефлексов. Таких, например, которые наблюдаются в экспериментах с цыплятами: цыплята избегают пищевых единиц характерного вида, если эта пища вызывала реакции отвращения у их сородичей. На самом деле здесь речь идет о т.н. социальной передаче избегания у цыплят. Эффект выглядит так. Новорожденным цыплятам

предъявляют бусину, смоченную жгучим веществом. Поскольку бусина оказывается жгучей на вкус, клюнувшие ее цыплята демонстрируют аверсивную поведенческую реакцию и в дальнейшем, естественно, отказываются клевать такие бусины. Однако замечено, что если другой цыпленок наблюдает процесс обучения, то далее он также начинает избегать клевания таких бусин. У цыпленка-наблюдателя образуется рефлекс. Аналогичные эффекты наблюдаются и у мышей, см., например, [Ивашкина и др., 2019].

Рассмотри схему организации этого эффекта. Фактически, здесь речь идет о некоторой модификации известной схемы формирования условного рефлекса. Когда контрагент клюет бусину, происходит оценка результатов этого действия. Результат действия оценивается и, в зависимости от него, изменяется величина связи между стимулом (бусина s_i) и действием (клевание, a_j). Например, в рамках автоматных моделей М.Цетлина меняется вероятность перехода p_{ij}: при положительной оценке (поощрение) p_{ij} увеличивается, при негативной (наказание) p_{ij} уменьшается [Цетлин, 1969].

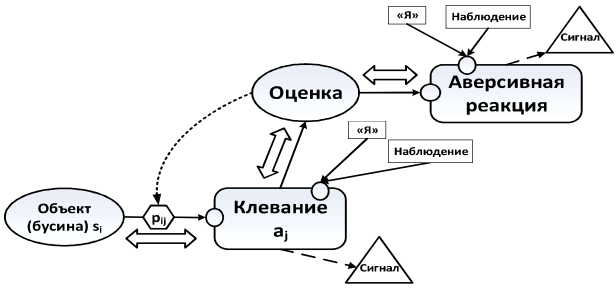

Рис. 7. Формирование реакции избегания. Двойными стрелками показано, что связь между компонентами является фактически двунаправленной

Введение явной компоненты С.Я, а также компоненты "Распознавание" позволяют описать процесс формирования такого же рефлекса и у наблюдателя. Это происходит из-за фактической замены сигнала "Я" наблюдаемым образом. Наблюдение за аверсивной реакцией контрагента в случае его отождествления с С.Я приводит к активизации вершины "Оценка", которая, в свою очередь, изменит вероятность p_{ij} (силу связи) между стимулом и действием. Т.е. образуется аналогичный рефлекс и у наблюдателя. Вторым важным моментом является необходимость признания наличия двунаправленных (взаимных) связей между компонентами. Схема на Рис. 7 – это типичная "функциональная" схема, описывающая направления движение информации, так, как это делается в классической теории управления. Однако переносить такой подход на

процессы, проистекающие в нервной системе, не совсем корректно: связи между нейронами или ансамблями нейронов имеют двунаправленный характер.

Такой подход к формированию социального рефлекса объясняет, в частности, разницу между собственным опытом и тем, который был приобретен в результате обучения (обучение на ошибках других). Собственный опыт является более устойчивым, прочным и продолжительным во времени. Чужой же забывается гораздо быстрее. Это объясняется тем, что активность компонент "Я" и "Наблюдение" разная.

Здесь важно отметить, что речь идет поведении, отличном от подражательного и тем более – контагиозного. Отличие это заключается в том, что здесь фигурирует компонента "Оценка", отсутствующая в схеме подражательного поведения.

3. Вопросы этики поведения

Есть все основания полагать, что наличие С.Я., как явной компоненты, понимание сути процессов отождествления С.Я. с наблюдаемыми контрагентами, а также наличие механизма подражания, – все это открывает путь к пониманию понятия морали поведения агентов с конструктивной точки зрения.

Следуя [Гусейнов, Апресян, 2000], будем считать, что слова "этика", "мораль", "нравственность" будут употребляться как взаимозаменяемые. Разумеется, моральная философия не рассматривает мораль, лишь как механизм групповой и индивидуальной адаптации посредством социальной организации поведения [Апресян, 2017]. Такой "механистический" подход более присущ эволюционным и этологическим теориям, однако, решая сугубо технические задачи, мы вынуждены придерживаться именно "адаптивного" понимания сути морали. При этом мы допускаем применимость моральных оценок к поведению искусственных агентов.

Далее определим три основных вопроса, определяющих содержание морального поведения: (1) зачем такое поведение нужно, (2) какова целевая функция (или основной регулятив), определяющая поведение, (3) каковы механизмы, лежащие в его основе.

1. Необходимость морали. Уже говорилось, что мораль – это механизм адаптации, то, что позволяет функционировать социуму более эффективно. При этом важнейшей особенностью морали является ее гибкость, вариативность. Эта надстройка над базовыми моделями поведения чрезвычайно "легковесна" и может варьироваться в широких пределах на протяжении жизненного цикла индивида.

2. Целевая функция регулирования поведения. Основополагающим регулятивом в межличностных отношений является т.н. "золотое правило" морали. Суть этого правила в определении отношений, основанных на практике взаимности ("ты – мне, я – тебе"). Генезис этого правила описан в работе [Апресян, 2013] и является отдельной темой. Здесь важно лишь отметить, что, во-первых, это правило задает целевую функцию морального поведения. А во-вторых, золотое 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) форме. Для реализации этого основного правила морального поведения требуется определение того, что такое "хорошо" и что такое "плохо" для индивида. Ответ на этот вопрос кроется в понимании сути механизмов, определяющих основу морального поведения.

3. Механизмы морального поведения. Очевидно, что подражательное поведение и социальное обучение можно рассматривать лишь как очень далекий базис для поведения, оцениваемого с точки зрения морали. Однако существует механизм, имеющий более явную связь с этикой поведения и называющийся эмпатией.

Эмпатия. Под этим термином понимается способность к отзывчивости на эмоциональные состояния окружающих индивидов разной степени близости. Разумеется, предрасположенность индивида эмпатии является необходимым, но не достаточным свойством для моральности индивида. Но в этом вопросе нас интересуют два аспекта эмпатии – механизм ее реализации, а также объект эмпатии.

Реализация механизма эмпатии возможна на том же принципе отождествления (или определения степени близости) наблюдаемого агента с С.Я. С объектом эмпатии, как отзывчивости на эмоциональное состояние, дело обстоит несколько сложнее. В определенном смысле эмоции – это, прежде всего, способ интегральной оценки состояния индивида (см. Информационную теорию эмоций П.Симонова, [Симонов, 1982]). Роль эмоций, как фактора, стабилизирующего поведение, контрастирующего сенсорное восприятия и пр. для деятельности не только животных, но и искусственных агентов (роботов), описана, например, в [Karpov, 2014]. Здесь же важно, что эмпатия – это основа для более высокого уровня управления, связанного с целеполаганием и планированием. В терминах моральной философии это означает действие золотого правила: либо реализовывать такой план действий, при котором контрагенту будет хорошо (увеличение уровня эмоционального состояния, позитивная формулировка правила), либо сформировать план, не ведущий к появлению отрицательных эмоций контрагента (негативная формулировка). В любом случае эмоциональное состояние контрагента оказывает влияние на формирование мотива поведения, его цели.

Таким образом, мы можем постулировать, что: (1) в перечень механизмов, определяющих нравственное поведение, наряду с прочими, входят социальное обучение и эмпатия и (2) основной мотивацией морального поведения (то, на что направлено золотое правило морали) является максимизация эмоционального уровня контрагента, на которого направлено воздействие индивида. Поскольку знак и величина эмоции непосредственно определяется существующими потребностями агента, то можно сделать весьма "механистический" вывод: действие основного морального регулятива направлено на удовлетворения потребностей индивида.

Такая неизбежная вульгаризация – это результат попытки переноса сугубо технических механизмов (вплоть до адаптационных моделей) на область, принципиально плохо формализуемую – моральную философию.

Заключение

Итак, было показано, что введение понятия С. Я. как конструктивного элемента системы управления позволяет предложить ряд моделей, реализующих такие важнейшие компоненты социального взаимодействия, как подражательное поведение и социальное обучение. Кроме того, эта компонента является конструктивной и в вопросе моделирования такой способности социального индивида, как эмпатия. Однако попытки механического переноса этих механизмов для объяснения или обоснования моральных принципов организации взаимодействия, даже как адаптационного механизма, логически приводят к весьма примитивным и банальным выводам. Отчасти причина тому – отсутствие общего понятийного аппарата.

В любом случае, если продолжать исследование аспектов моральности поведения аниматов "снизу", с "технической" стороны, то остается открытым целый ряд вопросов, таких, например, как:

1. Насколько нравственные правила определяются подражательным поведением, социальным обучением и способностью к эмпатии?
2. Насколько целесообразно интерпретировать формы поведения анимата именно с точки зрения основного нравственного правила?
3. Что подлежит регуляции в "моральном поведении"? Только ли характеристика близости "свой-чужой"?
4. Какова значимость фактора необходимости совместной деятельности для решения сложных задач в зависимости от свойств среды обитания и индивидуальных потребностей?

Эти вопросы требуют своего разрешения, причем вне зависимости от положений моральной философии, а оставаясь в рамках парадигмы моделей социального поведения роботов.

Список литературы

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

[Апресян, 2013] Апресян Р.Г. Генезис золотого правила // Вопросы философии. 2013. № 10. С. 39–49.

[Апресян, 2017] Апресян Р.Г. Этика: учебник. М.: КНОРУС, 2017. 356 с.

[Гусейнов, Апресян, 2000] Гусейнов А.А., Апресян Р.Г. Этика. М.: Гардарики, 2000. 472 с.

[Ениколопов, Кузнецова, Чудова, 2014] Ениколопов С.Н., Кузнецова Ю.М., Чудова Н.В. Агрессия в обыденной жизни. М.: Полит. энциклопедия, 2014. 496 с.

[Знаков, 2005] Знаков В.В. Психология понимания: Проблемы и перспективы. М.: «Институт психологии РАН», 2005. 448 с.

[Зорина, Полетаева, Резникова, 2002] Зорина З.А., Полетаева И.И., Резникова Ж.И. Основы этологии и генетики поведения. Учебник. 2-е изд. М.: Изд-во МГУ: «Высшая школа», 2002. 383 с.

[Ивашкина и др., 2019] Ивашкина О.И. и др. Социальная передача страха у мышей: влияние прошлого индивидуального обучения в задаче условно-рефлекторного замирания // Всероссийская с международным участием Конференция: XLIV Итоговая научная сессия «Системная организация физиологических функций». М., 2019.

[Карпов, Карпова, Кулинич, 2019] Карпов В.Э., Карпова И.П., Кулинич А.А. Социальные сообщества роботов. М.: УРСС, 2019. 352 с.

[Карпова, Карпов, 2018] Карпова И.П., Карпов В.Э. Агрессия в мире аниматов, или О некоторых механизмах управления агрессивным поведением в групповой робототехнике // Управление большими системами. 2018. Т. 76. С. 173–218.

[Кулинич, 2018] Кулинич А.А. Модель командного поведения агентов в качественной семиотической среде. Часть 2. Модели и алгоритмы формирования и функционирования команд агентов // Искусственный интеллект и принятие решений. 2018. № 1. С. 29–40.

[Лоренц, 1994] Лоренц К. Агрессия (так называемое «зло»). М.: Республика, 1994. 272 с.

[Лурия, 2007] Лурия А.Р. Лекции по общей психологии. Питер, 2007. 320 с.

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

[Резникова, 2004] Резникова Ж.И. Сравнительный анализ различных форм социального поведения у животных // Журнал общей биологии. 2004. Т. 65. № 2. С. 135–151.

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

[Тинберген, 1993] Тинберген Н. Социальное поведение животных. М.: Мир, 1993. 81 с.

[Цетлин, 1969] Цетлин М.Л. Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969. 316 с.