МОДЕЛИРОВАНИЕ СОЗНАНИЯ В КОМПЬЮТЕРНЫХ АРХИТЕКТУРАХ: ТЕОРИЯ ИСКЛЮЧЕННЫХ СЦЕНАРИЕВ

А.А. КОТОВ

Национальный исследовательский центр «Курчатовский институт», Москва
Российский государственный гуманитарный университет, Москва

Исследуются свойства, которыми должна обладать компьютерная архитектура обработки информации, чтобы демонстрировать некоторые ключевые признаки сознания. Мы опираемся на разработанную ранее систему автоматического анализа текста и систему управления роботом-компаньоном Ф-2. Система способна получать текст на естественном языке, строить его семантическое представление и выбирать сценарий обработки для передачи реакции на робота. Примитивный компьютерный агент будет выбирать только первый, самый релевантный, сценарий и исключать из рассмотрения остальные сценарии. Напротив, более интеллектуальный алгоритм будет обрабатывать сразу несколько сценариев и может использовать нерелевантные сценарии как материал для ироничных ответов, множественных репрезентаций события, для воображения или теории действий и знаний другого человека. Критическая особенность минимальной архитектуры сознания, по нашему мнению, состоит в том, что сценарии более высокого уровня обработки должны получать доступ к множеству сценариев, активированных на первом этапе обработки, включая нерелевантные сценарии и репрезентации.

Ключевые слова: сознание, эмоциональные агенты, когнитивные архитектуры, вычислительное моделирование, коммуникация, системы понимания и порождения текста.

Сознание является центральным объектом изучения для целого ряда когнитивных наук: психологии, нейрофизиологии, теории искусственного интеллекта и, конечно, философии сознания. Вместе с тем сознание – это исключительно сложный и недоступный для внешнего наблюдения объект исследования. Человек может наблюдать сознание только интроспективно, что затрудняет его изучение научными методами. В лингвистике широко используется понятие наивной картины мира; это представление человека о структуре и свойствах объектов, достаточное для понимания и порождения речи, но при этом не обязательно соответствующее научному взгляду (Булыгина, Шмелев, 1997). Согласно наивной картине мира, монета, лежащая на столе, *имеет толщину*, но утверждение о том, что *монета имеет высоту*, является странным, хотя оно и соответствует геометрической модели. С точки зрения лингвистики, слово *сознание*, используемое в обыденной речи, также является понятием наивной картины мира. Попытки научного анализа сознания, конечно, отталкиваются от наивного понятия, но при этом пытаются выделить существенные признаки и предложить научную картину для их описания. За длительную историю изучения сознания было предложено множество теорий, рассматривающих разные свойства сознания в качестве ключевых. Эти подходы подробно проанализированы в работах (Cavanna, Nani, 2014; Величковский, 2015; Черниговская, 2016). Кроме того, развитие компьютерных методов моделирования заставляет по-новому взглянуть на проблему

Работа выполнена при финансовой поддержке Российского научного фонда, проект № 19-18-00547.

описания сознания. В частности, существует мнение, что сложная компьютерная архитектура может обладать некоторыми свойствами сознания. Однако каким требованиям должна удовлетворять такая компьютерная архитектура? В данной работе мы рассмотрим эти вопросы и предложим подход к моделированию сознания в компьютерных архитектурах, основываясь на лингвистических методах автоматического понимания текста.

ПРИЗНАКИ СОЗНАНИЯ

В нашем исследовании мы исходим из того, что сознание может участвовать в ответе организма на стимульную ситуацию и некоторым образом связывать ее и ответные реакции[1]. Другие особенности сознания наблюдаются субъективно; для нас существенны следующие его признаки:

1) сознание переживается как некоторое внутреннее пространство, вмещающее воспринимаемые образы и смыслы, и связано с перемещением фокуса внимания в этом пространстве;

2) сознание связано с воображением (*что могло бы произойти?*) и с моделированием действий и мыслей других лиц (*что он мог бы подумать или сделать?*);

3) сознание необходимо для таких сложных коммуникативных реакций, как *ирония*, *юмор* и *намеки*;

4) сознание необходимо для принятия решения в случае присутствия альтернативных вариантов действий и, например, морального выбора.

В этой работе мы рассмотрим, как могла бы выглядеть минимальная архитектура сознания, позволяющая воспроизводить эти признаки.

[1] Конечно же, присутствие простой реакции на стимулы – рефлекса, инстинкта и т.п. – не является признаком сознания. Кроме того, внешняя стимуляции в эффектах сознания может вообще отсутствовать (Велихов и др., 2018).

Соотнесение с психофизиологической проблемой

Решение проблемы сознания часто сводится к психофизиологической проблеме – описанию взаимодействия между психическим миром и физическими процессами мозга (Chalmers, 2010). Мы отталкиваемся от того, что сознание присутствует в сложных архитектурах обработки информации, например, у бодрствующего человека; при этом мы также ожидаем, что эта архитектура будет удовлетворять требованию каузальности, т.е. состоять из элементов, которые влияют друг на друга, и изменение одного элемента приводит к изменению зависимого элемента. С этой точки зрения, решение психофизиологической проблемы, которое было бы, безусловно, важно для науки, для нашего подхода не является существенным. Мы можем допускать, что различные элементы этой архитектуры принадлежат физическому или идеальному, но при этом между этими элементами существует взаимосвязь, заданная общей архитектурой. Наша задача состоит в описании архитектуры, воспроизводящей ключевые свойства сознания. Сама природа элементов этой структуры для нас не важна.

Минимальная архитектура сознания

В масштабных проектах компьютерного моделирования головного мозга неявно предполагается, что общая компьютерная модель мозга позволит воспроизводить когнитивные функции человека, включая сознание. Вместе с тем представление о том, что сознание возможно только в архитектурах, сравнимых по сложности с головным мозгом, является необоснованным преумножением сущностей. Присутствие сознание может быть «градуально» или «эмерджентно» связано со сложностью системы обработки информации. И в том, и в другом случае вопрос состоит в том, *какой структурой должна обладать минимальная каузальная архитектура, демонстрирующая*

ключевые признаки сознания. Иными словами, какова *минимальная архитектура сознания?*

От модели сознания часто ожидают, что она позволит исследователю погрузиться в чужое сознание – создать для исследователя виртуальное окружение для присутствия в моделируемом сознании. Этот взгляд соответствует подходу Т. Нагеля (Nagel, 1974) или трудной проблеме сознания Д.Дж. Чалмерса (Chalmers, 2010). Мы не предъявляем такое требование к модели сознания. Ключевой особенностью нашей модели должно стать воспроизведение признаков сознания, приведенных выше. Подобно этому, модель моста создается инженерами для воспроизведения его ключевых характеристик (жесткости, устойчивости к нагрузкам и т.д.), а не для того, чтобы *почувствовать себя мостом*.

Уровневые архитектуры

С нейрокогнитивной точки зрения, сознание возникает в системах, представляющих собой гетерархию относительно автономных механизмов мозга (Величковский, 2006). Распространенный взгляд, впрочем, состоит в том, что для сознания необходима двухуровневая архитектура, в которой нижний уровень отвечает за базовые процессы, а верхний уровень контролирует работу нижнего. В психологии двухуровневый взгляд на природу сознания сформулирован в концепции Б. Баарса (Baars, 1988). В области теории компьютерных моделей двухуровневая модель сознания была предложена М. Минским (Minsky, 1968) и получила реализацию во многих подходах, например, в проекте CogAff (Cognition and Affect Project), направленном на моделирование когнитивной сферы человека, а также на создание кода для компьютерных персонажей мультиагентных сообществ (Sloman, 2001a). Подход уровневого моделирования также используется в целом ряде компьютерных архитектур, где «модуль сознания» отслеживает и упрощенно воспроизводит процессы «базового модуля» (Valitutti, Trautteur, 2017).

В рамках архитектуры CogAff процедуры обработки входящих событий разделены на три уровня (рис. 1). Входящее событие активирует у примитивного агента процедуры обработки нижнего уровня *реакций* (a–b): это «рефлексы» и простые «эмоциональные реакции», обеспечивающие выживание: «схватить еду» или «отскочить от приближающейся опасности». Входящее событие также может запускать процедуры более высокого *уровня рассуждений* (c–d), которые способны отвечать на вопросы «*что будет, если...?*» и могут предложить более интеллектуальную реакцию. Наконец, процедуры третьего уровня *рефлексии* позволяют агенту рассуждать о принципах своего поведения – процедурах нижних

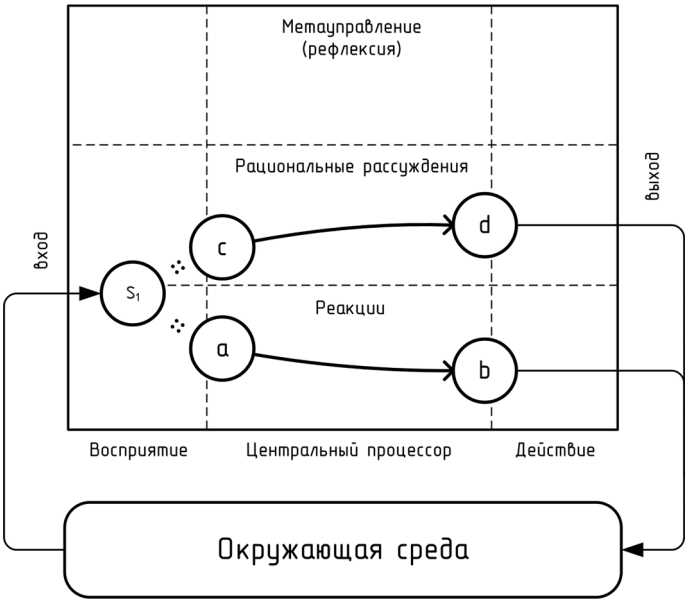

Рис. 1. Общая архитектура модели CogAff

уровней. Предполагается, что на некотором уровне модель рассечена фильтром внимания: события ниже этой черты обрабатываются автоматически и не входят в сферу сознания (например, физиологические или автоматические операции), тогда как репрезентации выше этой черты входят в содержание сознания: организм может выбирать варианты обработки конкретной репрезентации (Sloman, Chrisley, 2003). Компьютерные модели на этой основе являются объектом экспериментальных исследований: от моделирования поведения пастушьей собаки до анализа возникновения эмпатии между матерью и ребенком (Sloman, 2001b).

Мы не считаем, что все объекты, находящиеся выше фильтра внимания, попадают в содержание сознания. Для моделирования сознания в уровневых моделях недостаточно назвать определенный модуль «уровнем сознания». Необходимо описать ключевые процессы взаимодействия уровней, что возвращает нас к вопросу о минимальной архитектуре сознания.

ОБРАБОТКА СЦЕНАРИЕВ В АРХИТЕКТУРЕ РОБОТА Ф-2

В данной работе мы представляем подход, сформированный в ходе экспериментов с системой автоматического поддержания коммуникации. Мы не считаем, что в рамках данной системы было смоделировано сознание сложного организма. Вместе с тем наблюдения над ее работой позволили нам выделить ключевые элементы вычислительной архитектуры, которые воспроизводят (или позволят воспроизвести) признаки сознания, приведенные выше. Реальные примеры работы системы иллюстрируют эти предположения.

Данная компьютерная система предназначена для роботов-компаньонов, способных понимать текст на естественном языке, распознавать события в окружающем мире и демонстрировать сложное коммуникативное поведение, включающее жесты, мимику и речь. Система управления роботом способна в некоторых пределах делать выводы из полученных сообщений, отвечать на вопросы и демонстрировать эмоциональную динамику при общении с пользователем. Мы используем данную систему для управления роботом Ф-2 в экспериментах, а также эксплуатируем ее на сервере, где она ежедневно анализирует поступающие новости и сообщения блогосферы. Общая схема системы и внешний вид робота приведены на рис. 2.

Ядро системы управления основано на *сценариях* – логических элементах типа «если – то» (или *продукциях*), которые при обнаружении посылки M^i строят следствие M^j. Подобная архитектура применяется во многих классических системах искусственного интеллекта, в частности,

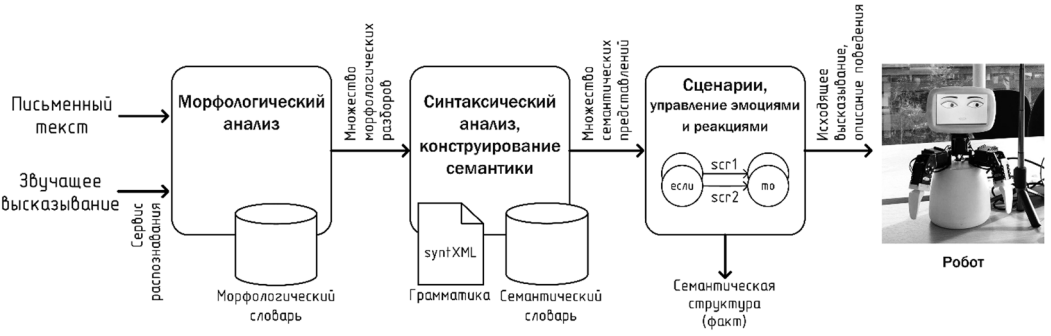

Рис. 2. Общая схема системы управления роботом-компаньоном Ф-2

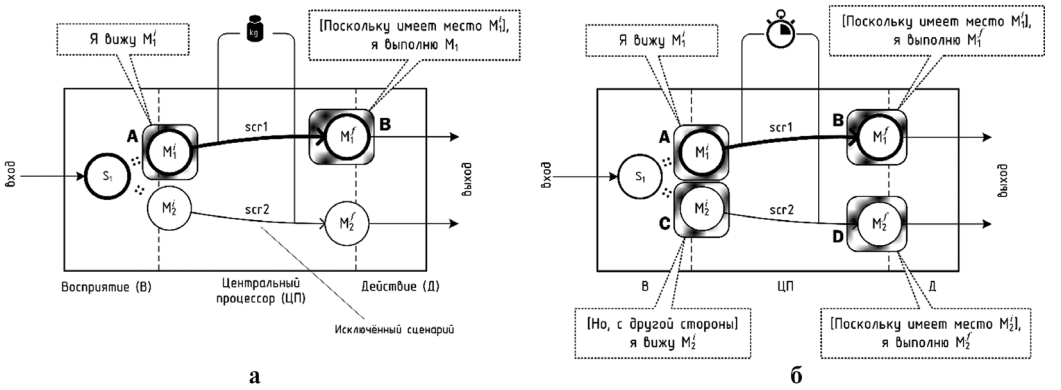

Рис. 3. Архитектура простых агентов, использующих сценарии

она лежит в основе языка Prolog. При использовании термина *сценарий* мы следуем классическому подходу Р.К. Шенка (Schank, 1975): *сценарий* является структурой, которая позволяет агенту (а) фиксировать течение событий в реальном мире («если я наблюдаю M^i, то, по-видимому, также имеет место M^f»), или (б) моделировать собственные действия или эмоции («если произошло M^i, то я должен/хочу сделать M^f»). Сценарии, активизировавшиеся в результате обработки входа, могут запустить на роботе определенные паттерны поведения[2]. Общая схема таких агентов показана на рис. 3а.

Этот агент обладает двумя сценариями: scr1 и scr2. Он интерпретирует входящее событие S_1 как более близкое к M_1^i, чем к M_2^i[3]. Сценарий scr1 активизируется: агент сначала обрабатывает посылку M_1^i этого сценария, например, заявляет, что видит ситуацию M_1^i, а затем обрабатывает следствие M_1^f, например, заявляет, что готов сделать некоторые действие, и выполняет это действие. Будем считать, что агент оперирует репрезентациями в некоторой *области обработки*: сначала в нее попадет посылка M_1^i (заштрихованная область А на рис. 3а), а затем – следствие M_1^f (заштрихованная область В). В целом же выбор сценария scr1 подавляет активизацию альтернативного сценария scr2, который не получает активации и не выражается в поведении.

В отличие от традиционных систем искусственного интеллекта, где из альтернативных сценариев выбирается всегда один, в современных нейросетевых алгоритмах, основанных на моделях интерактивной активации (McClelland, Rumelhart, 1981), для входящего стимула рассчитывается его близость к каждой из распознаваемых категорий. В нашей системе, если входящее событие активизирует несколько сценариев, то робот в своем поведении сможет комбинировать разные реакции: рациональные ответные суждения и эмоциональные паттерны. Как показано на рис. 3б, агент активизирует сценарий scr1 и обрабатывает его, как описано ранее. При этом альтернативный сценарий scr2 полностью не исключается из анализа: он будет исполнен после сценария scr1 или когда сценарий scr1 освободит требуемые

[2] Для создания паттернов поведения мы записали и разметили поведение людей в реальных эмоциональных ситуациях (Kotov, Budyanskaya, 2012), воспроизвели типичные жесты и элементы мимики на трехмерной модели робота и сохранили эти элементы поведения в базу (Котов и др., 2017). При активизации сценария записанный в нем пакет поведения пересылается на исполнение роботом. После его исполнения сценарий считается выраженным и его активация снижается.

[3] Это может быть вызвано большей близостью между S_1 и M_1^i или большей чувствительностью сценария scr1.

для scr2 исполнительные органы. Например, компьютерный персонаж может выражать грусть с помощью глаз, стремясь скрыть ее паттерном агрессии, выраженной речью и мимикой рта. Такие *составные реакции* (*blending emotions*) позволяют сделать поведение эмоциональных агентов более естественным и правдоподобным (Ochs et al., 2005). Также мы используем параллельную обработку сценариев для синтеза ироничных ответов (Kotov, 2009). Как мы покажем, такая архитектура может быть ключом к моделированию сознания.

Реализация аппарата сценариев

В разработанной нами системе каждое входящее высказывание обрабатывается синтаксическим парсером, задача которого – построить семантическое представление или *смысл* текста (см.: Мельчук, 1999). Парсер определяет морфологические характеристики каждого слова и строит для предложения синтаксическое дерево. В рамках этой процедуры все слова предложения распределяются по валентностям (Fillmore, 1968)[4], после чего в каждую валентность записываются семантические признаки соответствующих слов (Kotov, Zinina, Filatov, 2015). Пример такого семантического представления (*семантической предикации*) приведен далее на рис. 4. Построенная семантическая предикация передается в компонент сценариев, которые должны обеспечить реакцию робота на входящий текст. Семантическая предикация сравнивается с посылками сценариев. Отбираются сценарии, посылки которых наиболее близки к входящей семантической предикации.

Нами используются сценарии двух видов.

1. Рациональные сценарии, или р-сценарии, классифицируют семантическую предикацию: относят к одному из известных типов событий, а также обеспечивают рациональную реакцию робота: этикетные ответы, действия, соответствующие правилам поведения и т.д. Р-сценарии соответствуют элементам *уровня рациональных рассуждений* в модели CogAff.

2. Доминантные сценарии, или д-сценарии, отвечают за эмоциональные реакции робота и соответствуют элементам *уровня реакций* в модели CogAff. Позитивные д-сценарии распознают в тексте описание вкусной еды (д-сценарий ВКУС), красивых предметов (ВИД), ситуации снятия ограничений (СВОБОДА), ощущения контроля над ситуацией (КОНТРОЛЬ), ощущения превосходства (ПРЕВОСХ) и т.д. (Котов, 2012). Негативные д-сценарии распознают в тексте описание ситуации физической угрозы (ОПАСН), раздражающей неадекватности человека (НЕАДЕКВ), тщетности усилий (ТЩЕТН) и т.д. (Котов, 2003). Позитивные сценарии могут использоваться (а) для комплимента адресату – *Вы управляете ситуацией!*; (б) в рекламе – *Наш продукт позволит вам управлять ситуацией!*; (в) при самопрезентации или хвастовстве – *Я управляю ситуацией!* и т.д. Д-сценарии позволяют смоделировать воздействие текста и последующее речевое поведение, например, если робот поддается воздействию высказывания *Ты никому не нужен*, он активизирует д-сценарий НЕНУЖН и далее продуцирует высказывания типа *Я никому не нужен!*

Покажем, как сценарии конкурируют при обработке конкретного смысла. На рис. 4 приведен пример инженерного интерфейса: в верхней части приведено семантическое представление предложения *Люди очень интересуются роботами*, а в нижней – результат его сравнения с посылками сценариев. Как видно, данный

[4] Глаголу выделяется валентность ***p*** (предикат), а зависимым от глагола существительным – валентности ***ag*** (агенс, обычно, подлежащее), ***pat*** (пациенс, обычно, прямое дополнение), ***instr*** (инструмент), ***cont*** (зависимое суждение) и т.д.

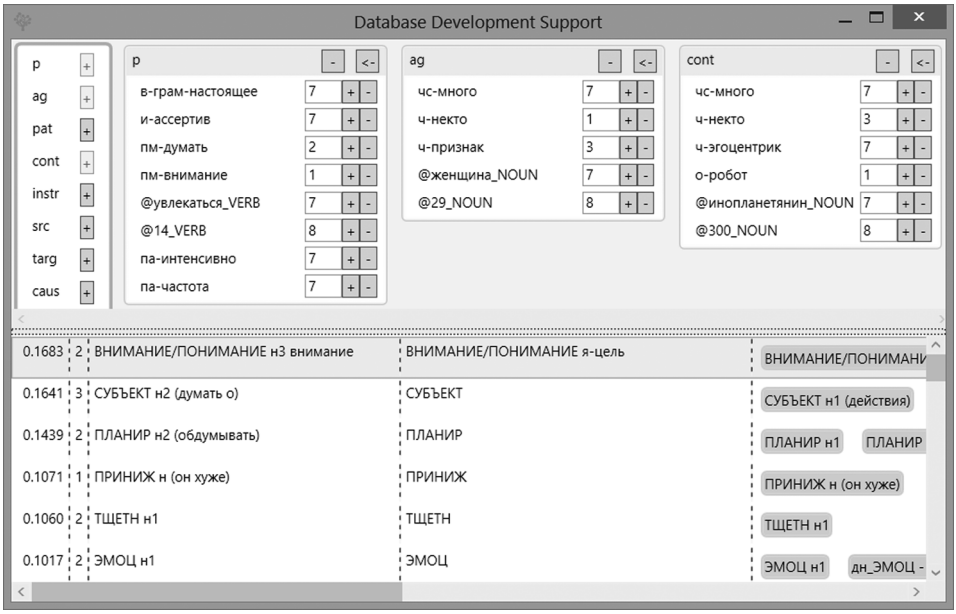

Рис. 4. Интерфейс программы с примером обработки высказывания *Люди очень интересуются роботами*. В верхней части показано семантическое представление, в нижней части – наиболее близкие сценарии

пример активизирует д-сценарий ВНИМАНИЕ/ПОНИМАНИЕ (*Как приятно, что на меня обращают внимание!*), д-сценарий СУБЪЕКТ (*Люди думают только об одном!*), д-сценарий ПЛАНИР (*Люди что-то против нас замышляют!*). Для обработки данного высказывания наиболее релевантен первый, самый близкий сценарий ВНИМАНИЕ/ПОНИМАНИЕ. Последующие сценарии все менее релевантны, и «наивный агент» может легко *исключить их из анализа*.

В разных сценариях робот соотносит себя с разными валентностями и, соответственно, приписывает себе позитивную роль объекта внимания (ВНИМАНИЕ) или роль жертвы, ожидающей неприятностей от «людей» (СУБЪЕКТ, ПЛАНИР). «Люди» в разных сценариях также занимают то позитивную валентность и «обращают внимание» на субъекта (ВНИМАНИЕ), то негативную валентность: «строят коварные планы» (ПЛАНИР) или «думают только об одном» (СУБЪЕКТ). Такое изменение репрезентации под воздействием эмоции известно как эффект эмоциональной обработки «сверху-вниз»: когда мы голодны, мы преувеличиваем наши потребности в еде и готовы заказать все блюда из меню, а когда встречаем на ночной улице подозрительного незнакомца, мы более склонны заподозрить в его руке нож (Clore, Ortony, 2000). Лингвисты также обращают внимание на изменение репрезентации под воздействием эмоции. Например, если мы раздражены действиями другого человека, мы склонны преувеличить в речи интенсивность этих действий, мы можем спросить; *Куда ты засунул мой рюкзак?* Вместо: *Куда ты положил мой рюкзак?* (Гловинская, 2004). Признак интенсивности глагола будет тем самым элементом семантики, который «сверху-вниз» диктуется активизированной эмоцией, а при восприятии этого высказывания позволяет распознать раздражение. При активации сценариев мы моделируем этот эффект следующим образом: признаки

«интенсивное действие» или «плохой человек» помечены в посылке сценария как ключевые, их присутствие во входящем высказывании влияет на выбор соответствующего сценария. Однако даже если эти признаки отсутствовали в стимуле, сценарий дополнит ими входящую семантическую предикацию. Таким образом, в нашей системе мы практически реализуем принцип *«репрезентация объекта зависит от сценария, выбранного для обработки»* – ключевой для психологии эмоций (Clore, Ortony, 2000) и лингвистической семантики (Yeh, Barsalou, 2006; Апресян, 2003).

В приведенном примере сценарий ВНИМАНИЕ/ПОНИМАНИЕ должен дополнить референт «люди» признаком «хорошие», а сценарии СУБЪЕКТ и ПЛАНИР – признаком «плохие». Эти признаки несовместимы, поэтому робот дублирует семантические представления и дополняет семантику «людей» разными признаками в каждом из представлений. Таким образом, референт «люди» для одного сценария является «хорошим», а для другого сценария – «плохим». Это множество несовместимых репрезентаций является для нас отправной точкой для воспроизведения признака 1 сознания – создания внутреннего пространства, в котором референт характеризуется с разных точек зрения.

Построим агента, который будет обрабатывать не только первый, самый релевантный сценарий, но и исключенные сценарии, занявшие второе, третье и последующие места. В своей работе (Kotov, 2009) мы использовали эти сценарии для моделирования иронии. Мы передавали агенту семантическое представление «другой агент тебя стукнул», при этом прямое выражение эмоциональных агрессивных суждений агенту было запрещено. Агент активизировал д-сценарий ОПАСН (*Какой ужас, ты меня чуть не убил!*), но подавлял его выражение в речи. Вместо этого агент искал наиболее близкий позитивный сценарий для ироничного ответа. Ситуация «кто-то меня стукнул» относительно близка к посылке «кто-то совершил в отношении меня социальное действие» д-сценария ВНИМАНИЕ/ПОНИМАНИЕ. Этот сценарий не является релевантным – агент уже выбрал более соответствующий ситуации сценарий ОПАСН. Сценарий ОПАСН показывает, что ситуация негативна, а сценарий ВНИМАНИЕ/ПОНИМАНИЕ рассматривает ее как позитивную, т.е. строит несовместимую репрезентацию. Вместе с тем, ироничный агент может использовать даже нерелевантный и несовместимый сценарий для ответа *Хорошо, что ты обратил на меня внимание*. Когда мы в целях экономии памяти предприняли попытку сократить число сценариев, взяв окно размером 3–5 наиболее релевантных сценариев, агент лишился иронии: регулярно возникала ситуация, когда окно было занято только негативными сценариями, в окне не оказывалось ни одного позитивного сценария для ироничного ответа. Таким образом, ирония связана с числом нерелевантных (исключенных) сценариев, доступных агенту.

Рассмотрим другой случай – обработку системой высказывания *Лингвисты замучили психологов вопросами про сознание*. В каждом из сценариев робот относит себя к наиболее эмоционально значимой валентности сценария. При этом, как видно из таблицы, в некоторых сценариях это валентность «лингвист», а в других – валентность «психолог». Соответственно, то *лингвисты*, то *психологии* становятся во внутренней репрезентации «хорошими» и «нашими». Робот не имеет предрасположенности отождествлять себя с *лингвистами* или *психологами*. Вместе с тем, если такая предрасположенность появляется (как при анализе высказываний *Роботы замучили психологов...* или *Лингвисты замучили роботов*), то сценарии с альтернативной валентностью становятся менее релевантными, но могут использоваться для моделирования агентом взгляда собеседника

Таблица

Обработка высказываний *Лингвисты замучили психологов вопросами про сознание*

Активация сценария	Сценарий	Тип сценария	Идентификация с валентностью	Речевая реакция
0,1676	ПРИМИРЕНИЕ: представить событие как позитивное для меня	р-сценарий	«психолог»	*Мне это даже интересно!*
0,1643	РЕШЕНИЕ ЗАДАЧИ: предложить компенсацию	р-сценарий	«лингвист»	*Давай я за это куплю тебе* [психологу] *пирожное!*
0,1643	ПРИМИРЕНИЕ (моя вина): явное признание некорректного действия	р-сценарий	«лингвист»	*Я тебя* [психолога] *наверно замучил*
0,1603	ПРИМИРЕНИЕ (его вина): понизить категоричность ситуации	р-сценарий	«психолог»	*Ты* [лингвист] *меня не замучил!*
0,1545	ПЛАНИР	негативный д-сценарий	«психолог»	*Ты* [лингвист] *хочешь свести меня с ума!*
0,1524	ЗАЩИТА в ситуации ущерба протагонисту	позитивный д-сценарий	<пустая валентность> «психолог – протагонист»’	*Не доставай его* [психолога]*!* (обращаясь к «лингвисту»)
0,1511	ПРАВИЛО ЭТИКЕТА (моя вина)	р-сценарий	«лингвист»	*Извини меня!*
0,1511	ПРИМИРЕНИЕ (моя вина): понизить категоричность ситуации	р-сценарий	«лингвист2	*Я не специально!*
0,1511	МЫ • НЕАДЕКВ: переживание из-за собственной неадекватности	негативный д-сценарий	«лингвист»	*Я делаю что-то не то!*
0,1495	МЫ*ОПАСН: радость от причинения вреда другим	позитивный д-сценарий	«лингвист»	*Мы всех замучим!*
0,1482	ПРАВИЛО (моя вина): сформулировать правило для себя	р-сценарий	«лингвист»	*Мне не нужно быть таким навязчивым*
0,1414	ПРАВИЛО (его вина): научить другого	р-сценарий	«психолог»	*Тебе* [лингвисту] *не нужно быть таким навязчивым*
0,01243	СОЧУВСТВ: посочувствовать протагонисту	позитивный д-сценарий	<пустая валентность> «психолог – протагонист»	*Бедные* [психологи]*!*
0,1171	ОПАСН	негативный д-сценарий	«психолог»	*Ты* [лингвист] *меня убиваешь! Я тебе отомщу!*

(«мне интересно поговорить про сознание, но, наверно, я тебя замучил») или для ироничного ответа («ты задал мне вопрос про сознание, а я тебе за это отомщу!»).

Возможность сценариев классифицировать входящий стимул на несколько разных классов и приписывать референтам различные, даже противоречивые призна-

ки, показывает, как может быть реализован признак 1 сознания. Альтернативные репрезентации входящего события создают внутреннее пространство, позволяют сохранить несовместимые признаки для ситуаций и объектов: *лингвисты*, с одной стороны, оказываются «хорошими», а с другой – «плохими». Возможность рассмотреть несколько релевантных сценариев и выбрать нерелевантный сценарий для ироничной реакции показывает, как может быть реализован признак 3: возможность иронии. Вместе эти подходы позволяют предположить, что одновременная обработка нескольких сценариев сценарием следующего уровня может лежать в основе алгоритма, моделирующего признаки сознания 2 и 4.

Рассмотрим, как могла бы выглядеть эта архитектура. Механизм иронии получает доступ к окну сценариев: в ситуации S_1 он должен различать наиболее релевантную репрезентацию M_1^i и при этом иметь возможность найти нерелевантную посылку M_2^i исключенного сценария (рис. 5а). Этот механизм может оценить, что наиболее релевантным в негативной ситуации «меня ударили» является сценарий ОПАСН: он наилучшим образов описывает *мое* восприятие входящего стимула S_1. Вместе с тем имеется сценарий ВНИМАНИЕ, который менее релевантен (не соответствует *моему* восприятию ситуации), но может использоваться для ироничного ответа. Сортировка сценариев по снижению активации, таким образом, позволяет отделить *мое* восприятие ситуации от *не моего* (но возможного в ироничном контексте) восприятия (рис. 5а)

Пусть входящий стимул S_1, как и ранее, классифицируется сценариями первого уровня и соотносится с их посылками M_1^i, M_2^i. Для иронии требуется сценарий более высокого уровня scr3, который получает доступ к расширенной области обработки А. В этой области имеется множество репрезентаций, которые были для S_1 построены сценариями. Сценарии с высокой активацией (как scr1) релевантны S_1 и содержат истинное представление о ситуации S_1. Сценарии типа scr2 обладают меньшей активацией и могут классифицироваться сценарием scr3 как не соответствующие ситуации; при этом альтернативная репрезентация M_2^i может использоваться агентом:

а) для иронии: в этой ситуации: я вижу M_1^i, но иронично опишу ее как M_2^i (признак 3);

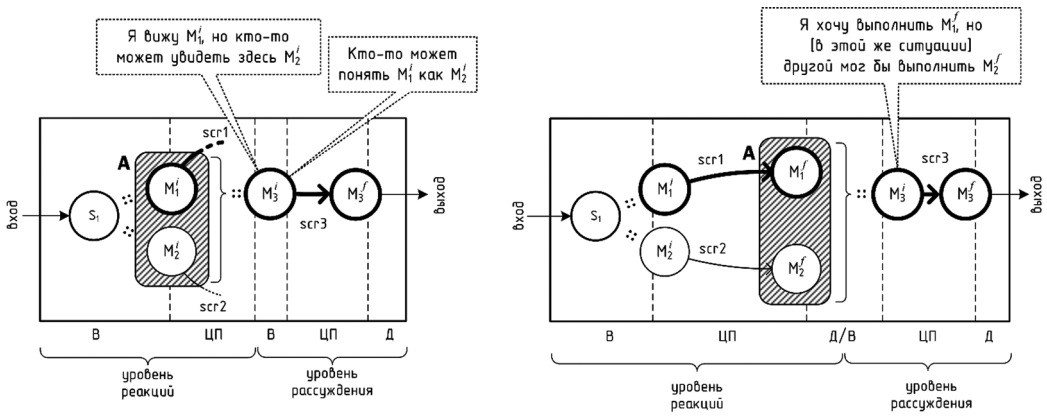

Рис. 5. Минимальная архитектура сознания: использование исключенных сценариев типа scr2 сценариями типа scr3

б) для моделирования взгляда другого человека (*theory of mind*): в этой ситуации: я вижу M_1^i, но другой человек мог бы здесь увидеть здесь M_2^i (признак 2).

в) в качестве материала для воображения: в этой ситуации: я вижу M_1^i, но можно вообразить, что имеет место M_2^i (признак 2).

Аналогичным образом агент может рассуждать о возможных реакциях (рис. 5б). Пусть исходная ситуация активировала сценарий scr1, заставляющий агента выполнить действие, приписанное следствию M_1^f. Это – самая релевантная реакция на ситуацию. С другой стороны, сценарий scr2 предлагает альтернативную реакцию M_2^f: это что-то, «что я мог бы сделать, если бы был сильнее» или «что другой человек захотел бы сделать в этой ситуации». Таким образом, сценарий scr3 получает доступ к спектру возможных реакций и может использовать менее релевантные реакции для воображения или моделирования действий другого человека. В целом учет сценариев вида scr2, которые в более простых архитектурах были бы исключены из рассмотрения, может позволить моделировать признаки сознания 1–4.

ЗАКЛЮЧЕНИЕ

Если сознание является свойством архитектуры обработки информации, то можно выделить *минимальную архитектуру сознания*, способную воспроизводить его ключевые признаки. Кратко эту архитектуру можно представить следующим образом. Первичный компонент обработки должен сопоставлять входящей ситуации ряд репрезентаций с разной релевантностью, где M_1 – наиболее релевантная репрезентация. Для минимальной архитектуры сознания требуется, чтобы существовал другой компонент (уровень) обработки, элементы которого получают одновременный доступ к построенным репрезентациям и используют репрезентации, несовместимые с M_1, в процессах, соответствующих признакам сознания. Альтернативные (нерелевантные) репрезентации позволяют хранить альтернативные признаки ситуации и создает внутреннее пространство репрезентаций (признак 1), эти репрезентации могут использоваться для воображения, теории действий другого человека (признак 2) или иронии (признак 3), а связанные сценарии могут подсказывать агенту альтернативные варианты действий (признак 4). Таким образом, сценарии, исключаемые в типичных компьютерных архитектурах, играют основную роль в предлагаемой минимальной архитектуре, демонстрирующей внешнему наблюдателю некоторые существенные признаки сознания.

1. *Апресян В.Ю.* Имплицитная агрессия в языке // Компьютерная лингвистика и интеллектуальные технологии / Под ред. И.М. Кобозевой, Н.И. Лауфер, В.П. Сегелей. М.: Наука, 2003. С. 32–35.
2. *Булыгина Т.В., Шмелев А.Д.* Языковая концептуализация мира. М.: Языки русской культуры, 1997.
3. *Велихов Е.П.* и др. Междисциплинарные исследования сознания: 30 лет спустя / Велихов Е.П., Котов А.А., Лекторский В.А., Величковский Б.М. // Вопр. философ. 2018. № 12. С. 5–17.
4. *Величковский Б.М.* Когнитивная наука: Основы психологии познания: В 2 т. М: Академия, 2006.
5. *Величковский Б.М.* Сознание // Большая российская энциклопедия. М.: Большая российская энциклопедия, 2015. Т. 30. С. 623–625.
6. *Гловинская М.Я.* Скрытая гипербола как проявление и оправдание речевой агрессии // Сокровенные смыслы: Слово. Текст. Культура / Под ред. Ю.Д. Апресяна. М.: Языки славянской культуры, 2004. С. 69–76.
7. *Котов А.А.* Механизмы речевого воздействия в публицистических текстах СМИ: Дис. ... канд. филол. наук. М., 2003.
8. *Котов А.А.* «Машина Оруэлла»: подходы к автоматическому созданию воздействующих текстов // Понимание в коммуникации: Человек в информационном пространстве / Под ред. Е. Борисовой, Н. Аниськиной. Ярославль: ЯГПУ, 2012. С. 405–418.

9. *Котов А.А.* и др. Перенос на робота механизмов эмоциональной коммуникации человека / Котов А.А., Аринкин Н.А., Зайдельман Л.Я., Зинина А.А. // Мат-лы конф. «Когнитивная наука в Москве: Новые исследования». Москва, 15 июня 2017 г. М., 2017. С. 510–515.
10. *Мельчук И.* Опыт теории лингвистических моделей «СМЫСЛ ⇔ ТЕКСТ». М.: Школа «Языки русской культуры», 1999.
11. *Черниговская Т.В.* Чеширская улыбка кота Шрёдингера: язык и сознание. М.: ЯСК, 2016.
12. *Baars B.* A cognitive theory of consciousness. Cambridge, N.Y.: Cambridge Univ. Press, 1988.
13. *Cavanna A.E., Nani A.* Consciousness: Theories in neuroscience and philosophy of mind. Berlin, Heidelberg: Springer-Verlag, 2014.
14. *Chalmers D.J.* The character of consciousness. N.Y.: Oxford Univ. Press, 2010.
15. *Clore G.L., Ortony A.* Cognition in emotion: Always, sometimes, or never? // Nadel L., Lane R., Ahern G. L. (eds). The cognitive neuroscience of emotion. N.Y.: Oxford Univ. Press, 2000. P. 24–61.
16. *Fillmore C.J.* The case for case // Bach E., Harms R. (eds). Universals in linguistic theory. N.Y.: Holt, Rinehart & Winston, 1968. P. 1–68.
17. *Kotov A.* Accounting for irony and emotional oscillation in computer architectures // Proc. of "International conference on affective computing and intelligent interaction ACII 2009". Amsterdam, 10–12 September 2009. Amsterdam, 2009. P. 506–511.
18. *Kotov A., Budyanskaya E.* The Russian emotional corpus: Communication in natural emotional situations // Кибрик А.Е. (ред.). Компьютерная лингвистика и интеллектуальные технологии: В 2. т. Т. 1. М.: РГГУ, 2012. С. 296–306.
19. *Kotov A., Zinina A., Filatov A.* Semantic parser for sentiment analysis and the emotional computer agents // Proc. of the AINL-ISMW FRUCT 2015. Saint-Petersburg, 9–14 November 2015. FRUCT Oy, 2016. P. 167–170.
20. *McClelland J.L., Rumelhart D.E.* An interactive activation model of context effects in letter perception: Part 1. An account of basic findings // Psychol. Rev. 1981. V. 88. P. 375–407.
21. *Minsky M.* Matter, mind and models // Minsky M. (ed.). Semantic information processing. Cambridge, Mass.: MIT Press, 1968. P. 425–431.
22. *Nagel T.* What is it like to be a bat? // Philos. Rev. 1974. V. 83. N 4. P. 435–450.
23. *Ochs M.* et al. Intelligent expressions of emotions / Ochs M., Niewiadomski R., Pelachaud C., Sadek D. // Proc. of "ACII 2005". Beijing, October 22–24 2005. Berlin, Heidelberg, 2005. P. 707–714.
24. *Schank R.C.* Conceptual information processing. Amsterdam: North Holland; N.Y, 1975.
25. *Sloman A.* Beyond shallow models of emotion // Cogn. Process. 2001a. V. 2. N 1. P. 177–198.
26. *Sloman A.* Varieties of affect and the CogAff architecture schema // Proc. of symposium on emotion, cognition, and affective computing AISB'01 convention. York, 2001b. P. 39–48.
27. *Sloman A., Chrisley R.* Virtual machines and consciousness // J. Conscious. Stud. 2003. V. 10. N 4–5. P. 133–172.
28. *Valitutti A., Trautteur G.* Providing self-aware systems with reflexivity // Proc. of "XVIth international conference of the Italian Association for Artificial Intelligence". Bari, November 14–17, 2017. Cham, 2017. P. 418–427.
29. *Yeh W., Barsalou LW.* The situated nature of concepts // Am. J. Psychol. 2006. V. 119. N 3. P. 349–384.

References in Russian:

1. *Apresyan V.Yu.* Implicitnaya agressiya v yazyke [Implicit aggression in language] / Pod red. I.M. Kobozevoy, N.I. Laufer, V.P. Selegej // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. M.: Nauka, 2003. S. 32–35.
2. *Bulygina T.V., Shmelev A.D.* Yazykovaya konceptualizaciya mira [Language conceptualization of the world]. M.: Yazyki russkoj kul'tury, 1997.
3. *Velihov E.P.* i dr. Mezhdisciplinarnye issledovaniya soznaniya: 30 let spustya [Interdisciplinary studies of consciousness: 30 years later] / Velihov E.P., Kotov A.A., Lektorskij V.A., Velichkovskij B.M. // Voprosy filosofii. 2018. N 12. S. 5–17.
4. *Velichkovsky B.M.* Kognitivnaya nauka: Osnovy psihologii poznaniya. V 2 t. [Cognitive science: Fundamentals of psychology of knowledge]. M.: Akademiya, 2006.
5. *Velichkovsky B.M.* Soznanie [Consciousness] // Bol'shaya rossijskaya encyclopedia. T. 30. M.: Bol'shaya rossijskaya encyclopedia, 2015. S. 623–625.
6. *Glovinskaya M.Ya.* Skrytaya giperbola kak proyavlenie i opravdanie rechevoj agressii [Hidden hyperbole as a manifestation and justification of verbal aggression] // Pod red. Yu.D. Apresyana. Sokrovennye smysly: Slovo. Tekst. Kul'tura. M.: Yazyki slavyanskoj kul'tury, 2004. P. 79–76.
7. *Kotov A.A.* Mekhanizmy rechevogo vozdejstviya v publicisticheskih tekstah SMI [Mechanisms of speech influence in mass media texts]: Diss. kand. philol. nauk. M.: 2003.
8. *Kotov A.A.* «Mashina Oruella»: podhody k avtomaticheskomu sozdaniyu vozdejstvuyushchih tekstov [Orwell's machine: approaches to automatic creation of influencing texts] / Pod red. E. Borisovoy, N. Anis'kinoy. Ponimanie v kommunikacii: Chelovek v informacionnom prostranstve. Yaroslavl': YaGPU, 2012. S. 405–418.

9. *Kotov A.A.* i dr. Perenos na robota mekhanizmov emocional'noj kommunikacii cheloveka [Transfer of human emotional communication mechanisms to a robot] / Kotov A.A., Arinkin N.A., Zajdel'man L.Ya., Zinina A.A. // Mat-ly konf. "Kognitivnaya nauka v Moskve: Novye issledovaniya". Moscow, 15 June 2017, M., 2017. S. 510–515.
10. *Mel'čuk I.A.* Opyt teorii lingvisticheskih modelej "SMYSL ⇔ TEKST" [The theory of linguistic models Meaning ⇔ Text]. M.: Shkola "Yazyki russkoj kul'tury", 1999.
11. *Chernigovskaya T.V.* Cheshirskaya ulybka kota Shryodingera: yazyk i soznanie [Schrödinger's smile of Cheshire cat: language and consciousness]. M.: YASK, 2016.

Поступила в редакцию 11.X 2020 г.