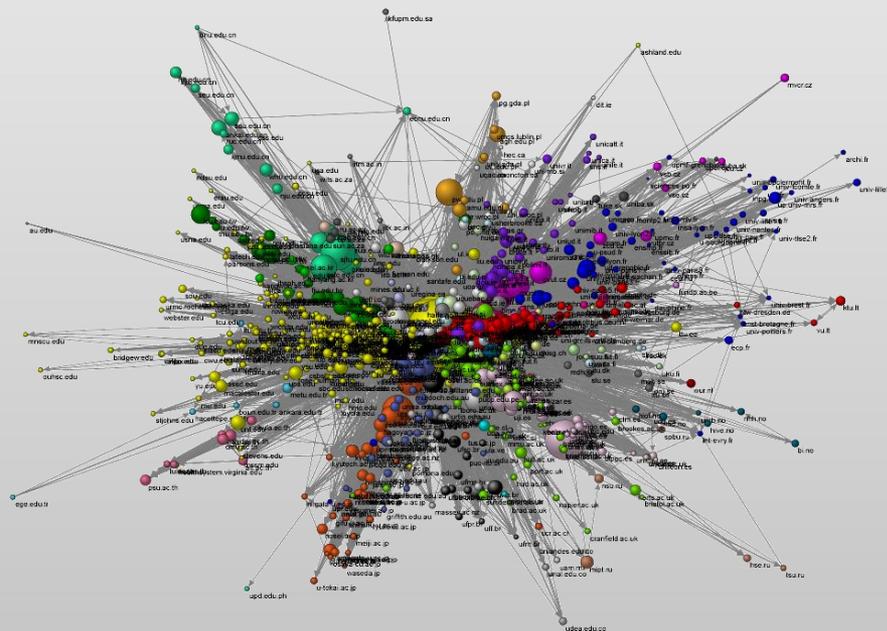


Проблема понимания и искусственный интеллект



О.П.Кузнецов

Институт проблем управления
им. В.А. Трапезникова РАН

Семинар «Этические проблемы искусственного интеллекта»
16.04.2025



ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
**ИНСТИТУТ
ПРОБЛЕМ
УПРАВЛЕНИЯ**
ИМ. В.А. ТРАПЕЗНИКОВА
РОССИЙСКОЙ АКАДЕМИИ НАУК

О ПОНИМАНИИ – ПРЕДВАРИТЕЛЬНО

Проблема понимания обсуждается с давних времен. Долгое время она считалась предметом гуманитарных наук: философии, психологии, лингвистики. Отдельно следует упомянуть когнитивную семантику, которая в последнее время считается разделом современной лингвистики, хотя в действительности когнитивной семантикой следовало бы считать исследование всех (не только языковых) средств возникновения, хранения смыслов и выражения их в языке, жестах, поведении и т.д.

Понимание — универсальная операция мышления, связанная с усвоением нового содержания, включением его в систему устоявшихся идей и представлений.

А.А. Ивин, А.Л. Никифоров. Словарь по логике/ М.: ГИЦ ВЛАДОС, 1997.

Википедия, 18.02.2025.

А. Алымова. Понимание. <https://blog.alter.ru/glossary/ponimanie/>

Результат понимания — собственная интерпретация чего-либо (текста, поведения, сновидений и т. д.). В этом контексте существует правильное и неправильное, глубокое и поверхностное, полное и неполное понимание.

О ПОНИМАНИИ – ПРЕДВАРИТЕЛЬНО

Знаков В.В. Психология понимания: проблемы и перспективы/ М.: Изд-во «Институт психологии РАН», 2005.

Чудова Н.В. Понимание – предмет исследования и объект моделирования // Искусственный интеллект и принятие решений. 2012, №4, С. 3-31.

Тарасов В.Б. Проблема понимания: настоящее и будущее искусственного интеллекта // Материалы V-й Международной научно-технической конференции (OSTIS-2015, Минск, БГУИР, 19-21 февраля 2015 г.). – Минск: Изд-во БГУИР, 2015. – С.25-42.

Осипов Г.С., Чудова Н.В., Панов А.И., Кузнецова Ю.М. Знаковая картина мира субъекта поведения. М.: Физматлит, 2018. 264 с.

ПОНИМАНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

В искусственном интеллекте важность проблемы понимания была осознана почти сразу после его возникновения (в 60-х – 70-х гг. XX века) в связи с работами по машинному переводу и разработкой вопросно-ответных систем. ИИ внес принципиально новый подход к этой проблеме, который можно назвать конструктивным, поскольку речь шла о том, чтобы создать программы, которые способны понимать.

Вот одно из первых определений «машинного понимания» (Ньюэлл, 1972):
система S понимает знание K, если она использует его всякий раз, когда K уместно.

Здесь понимание относится к знаниям и, следовательно, связано с использованием этих знаний для решения задач; соответственно, глубину понимания можно охарактеризовать тем классом задач, которые способна решать система.

ПОНИМАНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

В 70-80 гг. ведущим специалистом по созданию «понимающих» интеллектуальных систем был Терри Виноград (Terry Winograd). На основе его идей Г. Левеск (Levesque H.J.) в 2011 г. предложил тесты на понимание, которые он назвал схемами Винограда.

Типичный пример схемы Винограда:

Подарок не помещается в коричневый чемодан, так как он слишком мал. Какой из предметов слишком мал?

Замена всего лишь союза «так как» на «несмотря на то, что» меняет правильный ответ на противоположный:

Подарок не помещается в коричневый чемодан, несмотря на то, что он так мал. Какой из предметов слишком мал?

В 2021 г. лучшие результаты вопросно-ответных систем по ответам на схемы Винограда не превышали 62%.

ПОНИМАНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Со времени исследований Винограда прошло уже больше 40 лет. Ситуация с ИИ резко изменилась. ИИ демонстрирует огромные успехи и находится в центре внимания общества. Нам уже обещают (не обоснованно!) скорое появление человекоподобного ИИ. Чего же не хватает ИИ для того, чтобы стать «сильным»? Много чего, но в первую очередь понимания. Отсутствие понимания – один из главных аргументов в критике мифов о скором появлении человекоподобного ИИ. В популярных книгах про ИИ об этом много говорится. Мелани Митчелл (англ. 2019, рус. 2022) посвятила несколько глав своей книги «барьеру понимания» как важному препятствию на пути к сильному ИИ.



«Люди обладают глубоким и сущностным пониманием ситуаций, с которыми сталкиваются, но пока ни одна система ИИ не может с ними в этом сравниться. Хотя современные программы ИИ почти сравнялись с людьми (а в некоторых случаях и превзошли их) в выполнении ряда узких задач, у них нет понимания, которое помогает людям воспринимать окружающий мир, пользоваться языком и мыслить логически. Очевидным этот недостаток понимания делают нечеловеческие ошибки, совершаемые системами».

М. Митчелл. Идиот или гений? Как работает и на что способен искусственный интеллект

ПОНИМАНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Тест Винограда: мои вопросы к ChatGPT 26.02.2025

Подарок не помещается в коричневый чемодан, так как он слишком мал. Какой из предметов слишком мал?

В данном случае, "коричневый чемодан" является объектом, в который не помещается "подарок". Таким образом, предметом, который слишком мал, является **коричневый чемодан**.

Подарок не помещается в коричневый чемодан, несмотря на то, что он так мал. Какой из предметов слишком мал?

Судя по вашему описанию, кажется, что "коричневый чемодан" слишком мал для подарка. Возможно, это намек на то, что чемодан не может вместить даже маленький подарок. Если это загадка, то, возможно, ответ кроется в том, что чемодан не предназначен для хранения подарков вообще или он поврежден.

О ПОНИМАНИИ В НАУКЕ

В 2013 г. начал свою работу проект *Human Brain Project (HBP)*. Его инициатор и первый руководитель Генри Маркрам декларировал цель проекта как воссоздание мозга человека в самых мелких деталях.

Маркрам: *«От генетического, молекулярного уровня к нейронам и синапсам, далее к цепям нейронов, макроцепям, мезоцепям, долям мозга — до тех пор, пока не возникнет понимание того, как связаны между собой все эти уровни и как они определяют поведение и формируют сознание».*

В 2013 г. этот проект завершился. Его итоги экспертная комиссия, которая оценивала итоги проекта, охарактеризовала следующим образом:

HBP установил новую парадигму цифровой нейробиологии и новую междисциплинарную культуру сотрудничества. Среди особо важных достижений - ведущие цифровые атласы мозга, передовые платформы для моделирования мозга во всех масштабах, применение когнитивного моделирования и персонализированной медицины, а также выдающиеся достижения в области нейроморфных вычислений, нейро-инспирированной робототехники и искусственного интеллекта.

Понимания, как видим, не возникло.

ПОЧЕМУ HUMAN BRAIN PROJECT НЕ ДОСТИГ ОБЪЯВЛЕННЫХ ЦЕЛЕЙ?

Прежде всего, конечно, потому, что возникли колоссальные вычислительные и энергетические сложности. Однако существует одна принципиальная причина провала, которая остается, даже если оставить в стороне технологические проблемы. Она коренится в самой идее проекта.

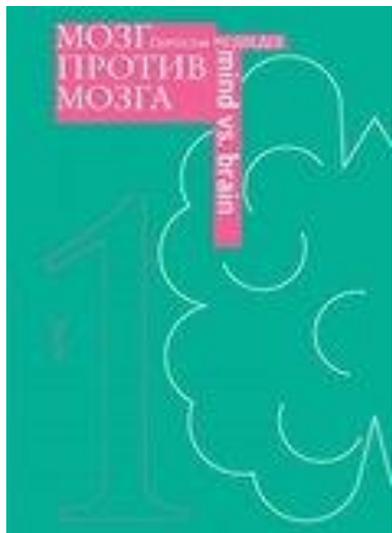
Еще раз процитируем Маркрама:

«От генетического, молекулярного уровня к нейронам и синапсам, далее к цепям нейронов, макроцепям, мезоцепям, долям мозга — до тех пор, пока не возникнет понимание того, как связаны между собой все эти уровни и как они определяют поведение и формируют сознание».

ПОЧЕМУ HUMAN BRAIN PROJECT НЕ ДОСТИГ ОБЪЯВЛЕННЫХ ЦЕЛЕЙ?

При подходе, предложенном Марккрамом, понимание не возникнет никогда, как бы велик и подробен не был объем полученных данных.

О ПОНИМАНИИ В НАУКЕ



*«Прорыв, произошедший в последние десятилетия в области технического оснащения исследований, как ни странно, не привел к кардинальному изменению ситуации в **понимании** работы мозга. Да, резко увеличилось число научных статей. Но в абсолютном большинстве эти работы сообщают нам о том, что в определенной зоне мозга обнаружена реакция на определенный раздражитель или что при выполнении определенной деятельности изменился рисунок ЭЭГ. Существенно меньше исследований, где показано, как именно происходит реорганизация работающих структур мозга при том или ином действии. ... По пальцам можно пересчитать статьи, где сообщается о механизмах работы мозга, какими законами он управляется, как взаимодействуют его структуры».*

С. Медведев. Мозг против мозга. М.: Бослен. 2017.

ОБ ИНТЕРПРЕТАЦИИ ДАННЫХ

Интерпретация данных не должна проводиться в терминах самих данных.

Для нее нужен язык более высокого уровня.

Это общий методологический закон для всех наук.

Частный случай этого закона – по структуре не всегда можно определить функцию. Функциональный язык – это, как правило, язык более высокого уровня. Он обычно является языком теории, интерпретирующим наблюдаемые структурные описания.

Эйнштейн: «Лишь теория решает, что мы ухитряемся наблюдать».

Вот, например, стандартная фраза в конце бланка результатов медицинского анализа, который мы получаем по электронной почте:

«Результат лабораторного исследования не является единственным параметром для постановки диагноза. Интерпретация результатов осуществляется только врачом».

МОЖЕТ ЛИ НЕЙРОБИОЛОГ ПОНЯТЬ МИКРОПРОЦЕССОР?

Прямой иллюстрацией к сформулированному выше тезису является эксперимент, описанный в статье

Jonas E, Kording KP (2017) Could a Neuroscientist Understand a Microprocessor? PLoS Comput Biol 13(1).

Нейробиологам была предложена задача: понять, как работает простейшая микропроцессорная система – чип MOS 6502, используемый в различных гаджетах, – используя методы, применяемые при экспериментальных исследованиях нервных систем.

Под микроскопом была изучена схема чипа. Измерялись напряжения на каждом соединении, определялось состояние каждого транзистора в разные моменты времени; отыскивались элементы схемы, отключение которых делало её неработоспособной, находились взаимные зависимости элементов и блоков, в частности, влияние каждого из них на работоспособность чипа в целом и т.п.

Однако **понять**, что делает этот микропроцессор, т.е. какова его функция, исследователям так и не удалось.

ЧТО ЗНАЧИТ «ПОНЯТЬ СИСТЕМУ»?

В этой же статье сформулирован четкий критерий понимания.

«Понимание части системы возникает, когда можно описать ее входы, выходы и преобразования от входов к выходам настолько точно, что эту часть можно заменить искусственным компонентом».

Добавим еще один вариант: «... настолько точно, что поведение этой части можно имитировать компьютерной программой».

Иначе говоря, нужна математическая модель этой части системы – именно математическая, а не квадратики со стрелками, – поскольку только математика умеет точно описывать преобразования.

Критерий математической корректности модели:

Описание модели должно позволить написать компьютерную программу, имитирующую ее поведение.

ЧТО ЗНАЧИТ «ПОНЯТЬ СИСТЕМУ»?



«Никакими экспериментами не объяснить, каким образом сотни миллиардов нейронов человеческого мозга посылают импульс в миг возникновения сознательного восприятия. Только с помощью математической теории мы поймем, каким образом психические процессы сводятся к нейронным импульсам. Нейробиологии требуются законы, аналогичные теории газов Максвелла-Больцмана, законы, которые объединяют между собой очень разные области».

S. Dehaene. Consciousness and the Brain. Deciphering How the Brain Codes Our Thoughts. Viking, New York, 2014.

С. Деан. Сознание и мозг. Как мозг кодирует мысли. М.: Карьера Пресс, 2018.

ПОНИМАНИЕ – СИЛЬНОЕ И СЛАБОЕ

Сформулированный выше критерий можно назвать критерием сильного, или полного понимания исследуемой системы. Сильное понимание – это понимание того, *как* устроена система.

Слабое понимание системы – это знание того, как будет реагировать система на те или иные входные воздействия. Такое понимание имеется у обычного пользователя бытового прибора, который умеет правильно нажимать кнопки, чтобы получать ожидаемые эффекты.

ОБЪЕКТИВНОЕ ПОНИМАНИЕ

Вспомним приведенное выше определение понимания:

Понимáние — универсальная операция мышления, связанная с усвоением нового содержания, включением его в систему устоявшихся идей и представлений. *Википедия*, 18.02.2025.

Переформулируем его следующим образом:

Понимáние — операция, связанная с усвоением нового содержания, включением его в существующую **картину мира**.

Но тогда возникает следующий вопрос: **где и в каком виде** она существует? И выясняется, что картин мира много. Существует общепринятое, устоявшееся научное знание о физическом мире. Его можно назвать *объективным*. Соответственно, можно говорить и об объективном понимании. Стремление понять, как работает мозг или чип, имеет в виду именно такое понимание.

Понимание в науке – 1) теория должна объяснить (интерпретировать) факты (*почему это произошло*); 2) теория должна предсказывать факты и события (*что будет, если*). Классический случай – после подстановки входных значений в формулы теории вычисление дает ответ, совпадающий с экспериментальными данными.

$$y = f(x) \quad I = \frac{U}{R}$$

СУБЪЕКТИВНОЕ ПОНИМАНИЕ И КОГНИТИВНАЯ СЕМАНТИКА

Если понятие объективной, научной картины мира возникло еще в конце XVIII века, то идея субъективной картины мира («Umwelt»), присущей каждому живому существу, впервые была высказана немецким биологом Якобом фон Иксюлем в 1909 г. Отметим, что кроме индивидуальных картин мира существуют и коллективные фрагменты картины мира, связанные как с так называемой «народной (интуитивной) физикой» (упрощенными моделями физического мира), так и с определенными социальными, религиозными, политическими представлениями о картине социального мира. Многие аспекты коллективных представлений изучаются в социальной психологии. Нас же будут интересовать субъективные, когнитивные аспекты картины мира человека, которые будем называть когнитивной семантикой.

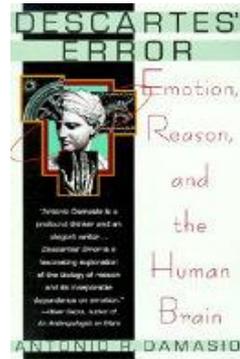
Для того, чтобы выяснить, почему создание понимающих интеллектуальных систем остается нерешенной проблемой, необходимо выяснить, как устроено и откуда берется понимание у человека.

НОВАЯ КОНЦЕПЦИЯ КОГНИТИВНОЙ СЕМАНТИКИ

1. Понимание – это интерпретация в терминах картины мира человека.
2. Картину мира строит наш мозг, и у каждого человека она своя. Эта картина (например, интуитивная физика: то, как человек представляет себе внешний мир) может сильно отличаться от картины мира других людей и от того, каким мир является на самом деле, т.е. каким его представляет научное знание.
3. Картина мира структурируется через категоризацию опыта человека, в ходе которой формируются понятия и связывающие их образно-схематические структуры.
4. Значения (смыслы) формируются раньше, чем формируются концептуальные структуры: они возникают из нашего допонятийного физического и социального опыта. Смыслы первичны, их языковое оформление вторично.
5. В основе смыслов лежат биологические цели и, в первую очередь, выживание.
6. В когнитивных процессах участвует не только мозг, но и тело.
7. Понимание связано с действиями в среде, знания о которой содержатся в картине мира.

НОВАЯ КОНЦЕПЦИЯ КОГНИТИВНОЙ СЕМАНТИКИ

В основе этой концепции лежат 4 книги:



Lakoff G. Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press, **1987**. Русский перевод: Лакофф Дж. Женщины, огонь и опасные вещи: что категории языка говорят нам о мышлении. М.: Гнозис, 2011.

Damasio, Antonio R. Descartes' error: emotion, reason, and the human brain / Putnam Publishing, 1994.

Pinker, S. The Stuff of Thought: Language as a Window in to Human Nature. New York, NY: Viking, **2007**. Русский перевод: Пинкер С. Субстанция мышления: Язык как окно в человеческую природу. М.: Книжный дом «ЛИБРОКОМ», 2013.

Seth Anil. Being You: A New Science of Consciousness. Faber and Faber. 2021. Русский перевод: Анил Сет. Быть собой. Новая теория сознания. М.: Альпина нон-фикшн, 2024.

НОВАЯ КОНЦЕПЦИЯ КОГНИТИВНОЙ СЕМАНТИКИ

1. Понимание – это интерпретация в терминах картины мира человека.
2. Картину мира строит наш мозг, и у каждого человека она своя. Эта картина (например, интуитивная физика: то, как человек представляет себе внешний мир) может сильно отличаться от картины мира других людей и от того, каким мир является на самом деле, т.е. каким его представляет научное знание. **Лакофф, Пинкер, Сет.**
3. Картина мира структурируется через категоризацию опыта человека, в ходе которой формируются понятия и связывающие их образно-схематические структуры. **Лакофф, Пинкер.**
4. Значения (смыслы) формируются раньше, чем формируются концептуальные структуры: они возникают из нашего допонятийного телесного опыта. Смыслы первичны, их языковое оформление вторично. **Лакофф, Пинкер.**
5. В основе значений лежат биологические цели и, в первую очередь, выживание. **Дамасио, Сет.**
6. В когнитивных процессах участвует не только мозг, но и тело. **Лакофф, Дамасио, Сет.**
7. Понимание связано с действиями в среде, знания о которой содержатся в картине мира. **Сет.**

1. ПОНИМАНИЕ И КАРТИНА МИРА ЧЕЛОВЕКА

1. Понимание – это интерпретация в терминах картины мира человека.

Интерпретация – это истолкование, осмысление. В разных областях это понятие конкретизируется по-разному. В математической логике оно определяется как соотнесение элементов формального языка с элементами некоторой предметной области. В искусстве под интерпретацией часто имеется в виду обнаружение нового смысла (классические примеры: «Кармен» у Мериме и Бизе, «Пиковая дама» у Пушкина и Чайковского).

В человеческом общении часто услышанное сильно отличается от сказанного: «ты меня не так понял» (неправильно интерпретировал).

Картина мира человека содержит только то, что ему понятно, осмысленно. Процесс понимания как интерпретация – это перевод на язык, в котором представлен нужный фрагмент картины мира.

К понятию понимания очень близко понятие **объяснения**. Можно сказать, что понимание – это объяснение самому себе. Объяснение другому – это перевод своих представлений на язык картины мира другого. Если картины мира (в той части, которая относится к предмету объяснения) совпадают или достаточно близки, объяснение, как правило, бывает успешным. В противном случае объяснить желаемое гораздо труднее; а иногда и вообще невозможно.

1. ПОНИМАНИЕ И КАРТИНА МИРА ЧЕЛОВЕКА

Следует отличать отсутствие понимания от неправильного понимания.

Отсутствие понимания означает, что воспринимаемое не удается вписать в существующую картину мира: у человека нет нужных понятий (и, соответственно, слов), чтобы его описать или задать разумные вопросы. В этом случае картину мира приходится достраивать, т.е. приобретать новые знания.

Неправильное понимание означает, что нужные понятия нашлись и интерпретация произошла, но она не соответствует реальному миру. При этом возможны два варианта.

Первый вариант: противоречие с картиной мира слишком явно; возникает когнитивный диссонанс и желание его устранить: либо интерпретация отвергается («Этого не может быть!», «Я не могу этого представить!»), либо картину мира приходится перестраивать (например, изменять свое мнение о каком-то знакомом человеке).

Второй вариант: явное противоречие сразу не обнаруживается, и человек долгое время (может быть, всю жизнь) с ним живет; например, считает, что Солнце вращается вокруг Земли или что все лебеди белые.

1. ПОНИМАНИЕ И КАРТИНА МИРА ЧЕЛОВЕКА

Человеку свойственно сохранять существующую картину мира и сопротивляться ее изменениям. Как отмечает создатель теории когнитивного диссонанса Леон Фестингер, при наличии двух альтернатив, одна из которых соответствует картине мира, а другая противоречит ей, *«человек будет искать такие источники информации, которые способствовали бы добавлению консонантных элементов, и будет избегать источников, увеличивающих диссонанс»*.

Перестройка картины мира требует значительных когнитивных усилий, часто сопровождаемых негативными эмоциями; поэтому стремление к ее сохранению согласуется с концепцией «когнитивной лени» Д. Канемана, т.е. минимизации когнитивных усилий.



2. КАРТИНА МИРА У КАЖДОГО СВОЯ



Нейронные коды сознания у каждого мозга свои. И хотя нейроны, отвечающие за кодирование цвета, формы или движения, у нас у всех одинаковы, складывающиеся из них структуры являются результатом долгого процесса развития, в ходе которого мозг приобретает уникальность, а действующие и бездействующие синапсы в сумме образуют нашу индивидуальность.

S. Dehaene. Consciousness and the Brain. Deciphering How the Brain Codes Our Thoughts. Viking, New York, 2014.

С. Деан. Сознание и мозг. Как мозг кодирует мысли. М.: Карьера Пресс, 2018.

2. КАРТИНА МИРА У КАЖДОГО СВОЯ

В основе картины мира человека лежат врожденные биологические инстинкты и потребности. Сама картина строится и перестраивается на протяжении всей жизни человека на основе его индивидуального опыта – как физического, так и социального. Как утверждает складывающаяся сейчас «новая теория сознания», перцептивный опыт определяется «определяется содержанием нисходящих предсказаний, а не восходящих сенсорных сигналов». Именно мозг, а не сетчатка, формирует наш зрительный мир. Мозг постоянно предсказывает свои ощущения и проверяет свои предсказания опытным путем, обучаясь на ошибках.

Это предиктивная обработка данных основана на ментальных моделях, существующих в мозгу и образующих картину мира. На психологическом уровне – это Система 1 Канемана.

2. КАРТИНА МИРА У КАЖДОГО СВОЯ

«Мы воспринимаем окружающий мир, чтобы эффективно в нем действовать, добиваться своих целей и в конечном итоге повышать свои шансы на выживание. Мы воспринимаем мир не таким, какой он есть, а таким, каким он нам полезен» (Сет).

Поэтому, например, картина физического мира (интуитивная, или «народная» физика) человека существенно отличается от картины, которую нам рисует научное знание. Уже законы Ньютона противоречат интуиции, не говоря о квантовой механике и теории относительности.

2. КАРТИНА МИРА У КАЖДОГО СВОЯ

Например, обычный слесарь-электрик незнаком с теорией электрических цепей и представляет электричество как поток некой жидкости, однако прекрасно справляется с ремонтом бытовых электроприборов и домашних электрических сетей. Это подтверждает приведенный выше тезис о том, что **обыденному мышлению нужна не истина, а польза.**

Истина не входит в число неотъемлемых жизненных целей человека. До нее, как правило, нужно докапываться, тратить когнитивные усилия, а действовать чаще всего нужно здесь и сейчас. Поэтому в стандартных ситуациях, составляющих большую часть жизни человека, действует быстрая Система 1 Д. Канемана, основанная на стереотипах. Но и медленная, рациональная Система 2 не всегда добирается до истины: либо из-за вычислительных трудностей, либо из-за недостаточных знаний, либо по причине когнитивной лени (Д. Канеман). Если же приходится выбирать между истиной и пользой, человек, как правило, выбирает (иногда вынужденно) пользу – вспомним историю Галилея, которому пришлось выбирать между истиной и выживанием.

2. КАРТИНА МИРА У КАЖДОГО СВОЯ

Тезис «не истина, а польза» относится прежде всего к обыденному мышлению. Научное знание претендует на истину, хотя для того, чтобы быть действительно истинным, знание должно быть доказательным. В естественных науках часть знания остается эмпирической, т.е. собранием фактов, для которых еще нет доказательной теории. Такова, например, наука о мозге. А вот в медицине доказательная часть пока невелика.

Эмпирические факты, не получившие объяснения в единой общепризнанной теории, становятся объектом различных интерпретаций и концепций. Так возникают различные научные школы, которые в нашем контексте можно было бы назвать групповыми картинами мира (точнее, некоторых его фрагментов).

2. КАРТИНА МИРА У КАЖДОГО СВОЯ

Таких групповых картин мира очень много в социальной сфере. Они могут быть связаны с религиозными или идеологическими воззрениями, иметь этническую, гендерную или классовую основу. Характерным примером групповой картины мира, основанной на ложном понимании, является карго-культ, возникший после второй мировой войны на островах Меланезии. Во время Второй мировой войны в ходе Тихоокеанской кампании США против Японии на этих островах американцы строили аэродромы и порты для доставки большого количества грузов, включая продукты питания, часть которых доставалась местным жителям. Когда война закончилась и грузы перестали поступать, аборигены решили, что эти грузы – дары небес, и если имитировать поведение американских летчиков и моряков, то дары небес появятся снова. Они делали ружья из дерева, самолеты из соломы, строили маяки, расчищали посадочные полосы и т.д. Но грузы так и не появились.

3. СТРУКТУРА КАРТИНЫ МИРА

Структура картины мира подробно описана в книге Лакоффа. Она имеет два уровня: а) базовый, «непосредственный», допонятийный, определяемый гештальтным восприятием и сенсорно-двигательным опытом; б) абстрактные понятийные структуры. Категории базового уровня представляют собой гештальты (целостно воспринимаемые образы); отношения между ними строятся на основе образных схем типа *вместилище, путь, связь, верх-низ, спереди-сзади, часть-целое, центр-периферия*, которые часто встречаются в нашем телесном опыте.

Со структур базового уровня, воспринимаемых непосредственно, начинается формирование картины мира у детей. Абстрактные понятийные структуры возникают либо в результате операций обобщения-конкретизации, либо с помощью метафорического переноса структур базового уровня на абстрактный уровень: так, например, категория времени характеризуется схемой спереди-сзади (будущее впереди, прошлое позади). При этом категории базового уровня находятся в середине иерархии общего-конкретного. Обобщение происходит вверх от базового уровня, конкретизация – вниз. Пример: собака – базовая категория, хищник – обобщение, овчарка – конкретизация. Эту структуру категорий впервые описала Элеонора Рош в своей теории прототипов.

3. СТРУКТУРА КАРТИНЫ МИРА

В рабочей (оперативной) памяти человека всегда присутствует только незначительная часть картины мира, и только к этой части предъявляются требования согласованности и непротиворечивости. В этом – одна из главных причин типичной для обыденного мышления фрагментарности и слабой чувствительности к противоречиям. Часто фрагмент картины мира, находящийся в рабочей памяти, называют репрезентацией.

Репрезентации – это конструкции, зависящие от обстоятельств. Они построены в конкретном контексте для конкретных целей: для осведомленности в данной ситуации, для того, чтобы быть готовым к требованиям текущей ситуации и понимать текст, инструкцию, проблему. Репрезентации учитывают всю совокупность элементов ситуации или задачи. Они специфичны, детализированы и непрочны. Репрезентация модифицируется, если изменилась вся ситуация или незаметный элемент вдруг стал заметным.

Знания – это конструкции, обладающие постоянством и существенно не зависящие от выполняемой задачи. Знания хранятся в долговременной памяти.

4. СЕМАНТИКА ПЕРВИЧНА, СИНТАКСИС ВТОРИЧЕН

Смыслы у человека возникают раньше, чем понятия и их воплощение в языке. Годовалые дети, еще не владеющие языком, уже имеют некоторый набор базовых значений, который они получают из своего телесного опыта. Только те понятия, которые «осмыслены» человеком, значимы для него, включаются в его картину мира, становятся его знанием. Осмыслить понятие для человека – это значит знать, как его можно использовать (опять – «польза, а не истина»).

Об этом говорит и упомянутое выше одно из первых определений понимания в ИИ:

система S понимает знание K, если она использует его всякий раз, когда K уместно.

4. СЕМАНТИКА ПЕРВИЧНА, СИНТАКСИС ВТОРИЧЕН

О вторичности синтаксиса говорит часто наблюдаемый факт: человек ищет слова, чтобы выразить свою мысль (мысль уже есть, а слов еще нет!). Вот что говорит А. Эйнштейн о том, как он думает:

«Слова, написанные или произнесенные, не играют, видимо, ни малейшей роли в механизме моего мышления. Психическими элементами мышления являются более или менее ясные знаки или образы, которые могут быть «по желанию» воспроизведены или скомбинированы. ... Элементы, о которых я только что говорил, у меня бывают обычно визуального или изредка двигательного типа. Слова или другие условные знаки приходится подыскивать (с трудом) только во вторичной стадии, когда эта игра ассоциаций дала некоторый результат, и может быть при желании воспроизведена».

4. СЕМАНТИКА ПЕРВИЧНА, СИНТАКСИС ВТОРИЧЕН

Человеку трудно передать дословно то, что он услышал 10 минут назад: он передает это приблизительно или вообще «своими словами», т.е. не то, что было сказано, а то, что он услышал, т.е. проинтерпретировал в своей картине мира. При этом разные люди одно и то же перескажут по-разному, - и не потому, что у кого-то из них хуже с памятью, а потому, что у них разные картины мира. Даже «суть сказанного» у них по этой причине может отличаться.

Кстати, о памяти – эти различия в пересказах услышанного или прочитанного говорят о том, что память хранит именно «суть», т.е. смысл, и не всегда слова. Отсюда вопрос, на который в когнитивной науке пока нет ответа: **как представлены смыслы в памяти человека?** Из вышесказанного видно, что, как правило, они представлены в более абстрактной форме, чем конкретные предложения, т.е. не синтаксически.

Схема Винограда – подарок и чемодан. Объяснить ответ, т.е. сконструировать логически правильное рассуждение, труднее, чем просто правильно ответить: потому что быстрый ответ основан не на логике, а на образной схеме «вместилище».

4. СЕМАНТИКА ПЕРВИЧНА, СИНТАКСИС ВТОРИЧЕН

В этой связи полезно вспомнить идею формализации, которая была предложена великим немецким математиком Д. Гильбертом для решения проблем математики. С подачи Тьюринга формализация лежит в основе всякой компьютеризации вообще.

Формализовать задачу – это значит представить ее как синтаксическую, лишенную всяких смысловых коннотаций и потому доступную для алгоритмической обработки, которая обеспечивает однозначный ответ (истина одна, хотя польз бывает много!).

5. ЗНАЧИМОСТЬ ОПРЕДЕЛЯЕТСЯ ЦЕЛЬЮ

Как возникают смыслы и значения?

Значимость воспринимаемых фактов и событий определяется их возможным влиянием на достижение тех или иных целей. Фундаментальной целью человека как биологического организма является выживание, т.е. поддержание жизненно важных параметров организма в допустимых пределах (избегание голода, холода, боли и т.д.). Для голодного человека особо значимой становится еда; если холодно, значима теплая одежда.

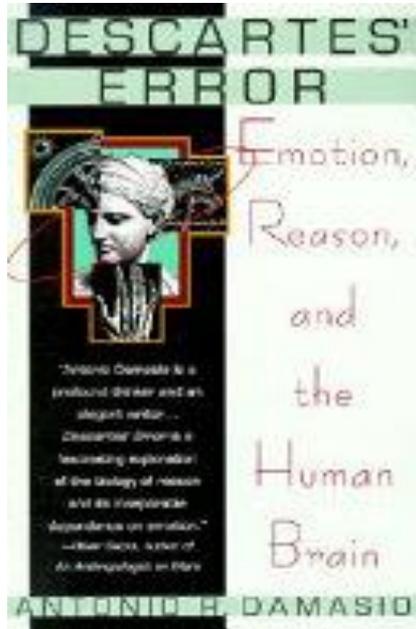
Очень часто выражение «бессмысленно» или «не имеет смысла» означает не «абсурдно», а «бесполезно»: не принесет пользы (опять польза!), т.е. не послужит достижению цели.

Опытные детективы и охотники-следопыты отличаются тем, что для них значимыми для целей поиска (**релевантными**, т.е. относящимися к делу) оказываются вещи, которые не замечаются (благодаря их кажущейся незначимости) непрофессионалами или новичками. **Релевантность** – это еще одно важное семантическое понятие, плохо формализуемое в ИИ.

5. ЗНАЧИМОСТЬ ОПРЕДЕЛЯЕТСЯ ЦЕЛЮ

Человек живет в обществе. Поэтому для него выживание – это не только поддержание физиологических параметров, совместимых с жизнью, но и улучшение таких социальных параметров, как качество жизни, «выживание в обществе», т.е. занятие в нем определенного места, обеспечивающего это качество, и т.д. Поэтому сложные социальные цели порождают соответствующие социальные смыслы и значения.

6. КОГНИТИВНЫЕ ПРОЦЕССЫ – НЕ ТОЛЬКО МОЗГ, НО И ТЕЛО



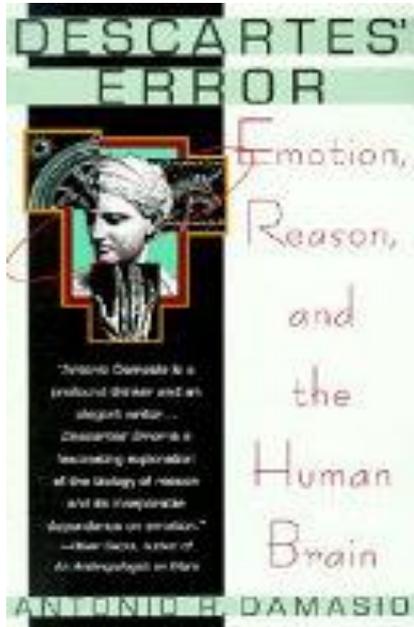
Телесные состояния и механизмы напрямую влияют на когнитивные процессы. Об этом говорят все авторы указанных 4 книг. А книга известного нейropsychолога А. Дамасио целиком посвящена обоснованию этого тезиса. Характерно уже ее название: «Ошибка Декарта» (известно, что Декарт резко отделял разум от тела).

Во-первых, тело является источником сведений о физическом мире и о своем внутреннем состоянии. Это очевидно и оспаривалось разве только сторонниками солипсизма.

Во-вторых, тело – это не только набор пассивных датчиков, но и участник процессов обработки данных, полученных от этих датчиков.

В третьих, тело – это биологический организм, представляющий собой целостность, поддерживающую свои жизненно важные параметры в определенных границах, и активно сопротивляющуюся выходу их за эти границы.

6. КОГНИТИВНЫЕ ПРОЦЕССЫ – НЕ ТОЛЬКО МОЗГ, НО И ТЕЛО



А. Дамасио выдвинул и обосновал концепцию *соматических маркеров* – процессов и ощущений в теле, связанных с эмоциями: кожно-гальваническая реакция, используемая в детекторах лжи, учащение сердцебиения, повышение кровяного давления и т.д. Он показал, что эмоции и их физические проявления – соматические маркеры – представляют собой грубую оценку различных вариантов (положительную или отрицательную). Тем самым они участвуют в процессе принятия решений в качестве подсказок, мобилизуя внимание и рабочую память.

Такие подсказки могут приводить к ошибкам; сильные эмоции могут «выбить человека из колеи»; понятно, что окончательное решение должно остаться за рациональным мышлением. Однако известны случаи, когда человек, лишенный эмоций в связи с какой-то патологией, вообще не способен принять решение.

7. ПОНИМАНИЕ СВЯЗАНО С ДЕЙСТВИЕМ

«Мы воспринимаем окружающий мир, чтобы эффективно в нем действовать, добиваться своих целей и в конечном итоге повышать свои шансы на выживание» (Сет).

Действия необходимы не только для достижения своих целей, но для корректировки картины мира. Мозг постоянно предсказывает свои ощущения и проверяет свои предсказания опытным путем, обучаясь на ошибках. Минимизацию ошибок предсказания путем действий Сет (с подачи Фристонa) называет **активным выводом**.

Действия не обязательно должны быть физическими. В социальной среде важную роль играет коммуникация: диалог, уточняющий картины мира собеседников; чтение лекции, сообщающей новое научное знание; пропаганда, навязывающая свою картину мира и т.д.

ВЫВОДЫ: ПОНИМАЮЩАЯ СИСТЕМА ДОЛЖНА БЫТЬ РОБОТОМ

Изложенная концепция приводит к неожиданному выводу: интеллектуальная система, способная понимать, – это **автономный активный агент (т.е. робот)**, действующий в среде, умеющий ставить цели, стремиться к их достижению и способный формировать и корректировать свою картину мира.

Любые системы обработки текстов никогда не будут понимающими: они не обладают активностью и потому не имеют собственной обратной связи.

Такой робот может быть не только материальным, но и виртуальным, т.е. быть программой, живущей в Интернете, содержание которого тем самым станет его окружающей средой.

Однако современные роботы не таковы. Они автономны при достижении целей, но эти цели им ставит человек. Их картины мира создаются программами, которые в них вложили люди. Любой современный робот может быть полностью перепрограммирован.

Это обстоятельство напрямую связано с **проблемой субъектности интеллектуальных систем.**

ВЫВОДЫ: РОБОТЫ И ПРОБЛЕМА СУБЪЕКТНОСТИ

Проблема субъектности активно обсуждается в рамках правовых аспектов, постоянно возникающих при эксплуатации интеллектуальных систем. Речь идет о том, можно ли считать автономную интеллектуальную систему субъектом права, т.е. может ли такая система отвечать за свои поступки?

Этот вопрос по существу равносителен вопросу: **можно ли считать робота личностью?**

Источник целей человека – он сам как биологическое и социальное существо. Он обладает неотъемлемыми чертами, определяющими его как личность. Часть этих черт – врожденная, формирующая его биологические цели, основанные на выживании; часть – приобретенная личным опытом; к ней относятся физические и социальные навыки и формируемая в течение всей жизни картина мира.

Поэтому **робот, которого можно полностью перепрограммировать, личностью считаться не может.**

ВЫВОДЫ: НУЖНО ЛИ СОЗДАВАТЬ ПОНИМАЮЩЕГО РОБОТА?

Получается, что понимающему роботу нужны свои собственные неотъемлемые цели, скажем, то же выживание, т.е. самосохранение. Но не появятся ли тогда у него свои «три закона робототехники», где на первом месте будет он сам, а не человек? И будет ли возникшее у него понимание похоже на человеческое? Парадокс – стремясь получить человекоподобный интеллект, мы рискуем получить нечто принципиально «человекоподобное». Нам это надо?

ЗАКЛЮЧЕНИЕ

Сет. «Нам не стоит слепо и бездумно добиваться стандартной цели ИИ – **воспроизвести**, а затем **превзойти** человеческий интеллект. Мы создаем разумные инструменты, а не коллег. Если мы действительно внедрим в мир новые разновидности субъективного опыта, нам придется иметь дело с нравственно-этическим кризисом беспрецедентных масштабов».

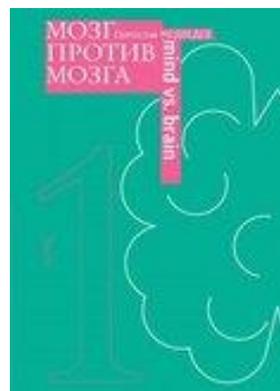
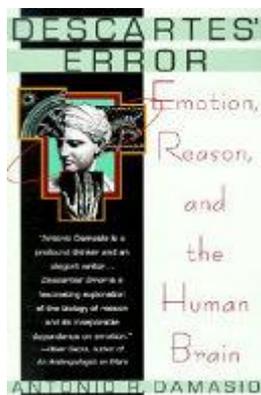
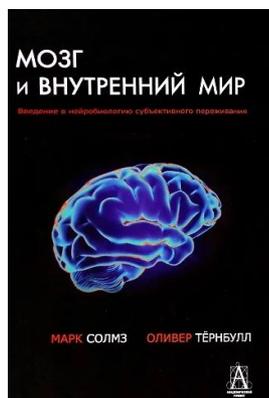
От себя добавлю:

«Воспроизвести» не удастся. Человеческий интеллект возможен только в человеке.

«Превайти» – не исключено, но лучше не надо.

Кузнецов О.П. О машинном обучении, мифах о сильном ИИ и о том, что такое понимание // Онтология проектирования. – 2024. – Т.14, №4(54). – С.466-482. – DOI: 10.18287/2223-9537-2024-14-4-466-482.

ЗАКЛЮЧЕНИЕ



СПАСИБО ЗА ВНИМАНИЕ!

Лучше потерять несколько лет, пробуя свои силы в чем-то принципиально новом, чем проводить отработанные эксперименты, которыми занимаются все кому не лень и которые кто-то другой может провести не хуже (если не лучше), чем вы.

Э.Кандель. В поисках памяти. М.: Астрель, 2012.