"Искусственный интеллект" как аргумент в споре о мышлении

Доклад на методологическом семинаре «Многомерный образ человека»

Игорь Михайлов

Москва, ИФРАН
06 декабря 2011

ИИ: история и теория

- Термин «искусственный интеллект» (ИИ) был введён Джоном Маккарти в 1956 г.
- С середины 1960-х работы по ИИ финансировались в Великобритании и США военными ведомствами.
- Примерно с 1974 г. в обеих странах было урезано финансирование проектов ИИ, не имеющих определённой прагматической цели, в связи с недостижением ожидавшихся результатов.

ИИ: история и теория

- 1980-е годы: возрождение интереса к ИИ по причине коммерческого успеха “экспертных систем”
- 1980-90-е гг.: взлёты и падения
- Начало 2000-х: новые академические исследования ИИ на фоне потребностей логистики, медицинской диагностики и некоторых других практических сфер

ИИ: история и теория

Среди направлений «академического ИИ», развиваемых в настоящее время:

- логическое программирование
- нейронные сети
- мультагентные системы

ИИ: история и теория

Постепенное исключение из этой области уже решённых проблем:

- «Например, в 1956 г. оптическое распознавание символов (ОРС) рассматривалось в качестве ИИ, а сегодня сложные ОРС-приложения с контекстно-зависимой проверкой орфографии и грамматики бесплатно поставляется вместе с большинством графических сканеров. Никто более не рассматривает сегодня уже решённые проблемы вычислительной науки, подобные ОРС, в качестве искусственного интеллекта» (http://en.wikipedia.org/wiki/AI#History)

Вокруг теста Тьюринга

1950 г. – Алан Тьюринг – принцип идентификации машинного интеллекта, который вошёл в историю как «тест Тьюринга» (ТТ).
Условия:
• человек (тестер) общается с другим человеком и машиной через клавиатуру и терминал;
• правила общения не заданы: может быть любая тема, любая продолжительность, любой язык, допускается использование сленга, ложь и т. п.;
• тестер должен определить, кто из собеседников является машиной; если он может идентифицировать машину с вероятностью, не превышающей простое статистическое распределение, машина считается прошедшей тест, а следовательно мыслящей.

Вокруг теста Тьюринга

Атаки на ТТ:

1. Соображения о логической связи теста и теоремы Гёделя о неполноте

- Теорема Гёделя справедлива для всех формальных систем, а компьютерные программы – это формальные системы.
- Следовательно, всегда будет хотя бы одна формула, истинность которой будет неразрешима для машины, но очевидна для человека.
- Следовательно, аналитически верно, что машина (или исполняемая в ней программа) не может быть моделью человеческого интеллекта.

Вокруг теста Тьюринга

2. Аргумент от qualia

Компоненты сознания, принципиально не доступные для компьютерного моделирования

(1) самосознание

(2) qualia

Вопрос о qualia:

- (условно) «бихевиористы»(Витгенштейн и последователи):
- традиционно ориентированные «философы сознания» (Сёрл, Перри)

Очевидно, что Тьюринг скорее склонялся к первой точке зрения.

Вокруг теста Тьюринга

Гипотеза систем физических символов (СФС)

(Ньюуел и Саймон,1976)

- Любая система физических символов (физических объектов, представляющих другие объекты в соответствии с некоторыми правилами именования) обладает необходимыми и достаточными средствами для сознательного действия.

Можно сказать, что позиция Деннета в его споре с Сёрлом в значительной мере основывается на идеологии СФС.

«Сильный ИИ» и «китайская комната»

Джон Сёрл:

- **«сильный [принцип] ИИ»** (концепции, сводимые к СФС)
  Любая программа, эффективно имитирующая интеллектуальные действия, и есть собственно интеллект.
- **«слабый [принцип] ИИ»**
  Для моделирования ИИ, возможно, понадобится определённый физический субстрат, искусственные нейронные сети, возможно, даже квантово-вероятностные эффекты, но всё же оно в принципе возможно.

«Сильный ИИ» и «китайская комната»

Аргумент «китайской комнаты»

Мысленный экперимент:

- человек, ни слова не понимающий по-китайски
- закрытое помещение, где в качестве интерфейса с внешним миром имеются щель для входящих записок и щель для исходящих записок
- некий справочник, в котором одним выражениям китайского языка сопоставлены другие по некоторым ему не известным правилам.

Сёрл: вся «китайская комната» как система способна пройти ТТ, хотя никто и ничто в ней не понимает китайского.

Метафора: компьютерная программа, умеющая оперировать символами, не содержит «интеллекта».

«Сильный ИИ» и «китайская комната»

Сёрл: концепция сознания как системы **интенциональных состояний**:

- по природе своей они аналогичны qualia в том смысле, что необходимо связаны с субстратом мозга и как таковые воспринимаются исключительно «в первом лице».

Мораль аргумента «китайской комнаты»: машина, основанная на СФС, не может иметь интенционального состояния, соответствующего пониманию.

«Сильный ИИ» и «китайская комната»

Сёрл здесь подобен карточному фокуснику:

- человек – для отвода глаз
- именно ему мы интуитивно приписываем понимание
- на самом деле в данной ситуации **«субъектом понимания» является... справочник**
- человеку отведена техническая роль информационного транспорта.

В том примитивном виде, как он описан Сёрлом, справочник вряд ли пройдет ТТ.

Если же справочник будет составлен более тонко, тогда почему бы и нет?

Всё дело в степени сложности программирования.

«Сильный ИИ» и «китайская комната»

Сильный аргумент Сёрля:

Если всю ситуацию поместить в голову человека – например, заставить его **выучить справочник наизусть**.

Понимания мы не добъемся и в этом случае: будучи спрошен **по-китайски**, понимает ли он китайский, человек **ответит «да»**, но будучи спрошен о том же **по-анлийскски**, он **ответит «нет»**.

«Сильный ИИ» и «китайская комната»

Возможный контраргумент:

Если в изначальной ситуации мы предположим, что человек всё же понимает китайский, но действует по условиям эксперимента: берёт записку, ищет выражение в справочнике, выписывает сопоставленное ему выражение...

- Что меняется для тех, кто снаружи?
- Становится ли китайская комната как целостный агент коммуникации более *понимающей*?

Субстанциализм и функционализм

[Переписка Деннета и Сёрла]

Главный аргумент Сёрла

- биохимический субстрат мозга обладает "достаточными каузальными силами" для того, чтобы причинно обусловить сознание = он необходим и достаточен для сознания.

Позиция Сёрля в основе своей имеет доктрину **субстанциализма**, а позиция Деннета может быть интерпретирована как **функционализм**.

Субстанциализм и функционализм

Деннет – аналогия с полётом:

- одно время **полёт так же считался функцией "биологического субстрата" птицы**, и это убеждение ничем не помогало братьям Райт
- пока они не подошли к вопросу с точки зрения законов аэродинамики и конечной цели строительства летательного аппарата, **отказавшись от попыток прямой имитации природы**.

Субстанциализм и функционализм

Контрвозражение Сёрла:

- он имел в виду не "секрецию" в буквальном смысле слова, а "каузальные силы", которые, по его мнению, могут содержаться не только в биологическом субстрате мозга, но любой его заменитель должен обладать эквивалентными "каузальными силами" для производства сознания.

Субстанциализм и функционализм

Что значит быть причинно зависимым от субстрата?

- Имитация природы в ИИ = создание «нейронных сетей» = те же компьютерные программы, только более сложные.

Тогда:

- загадочная причинная зависимость от субстрата = функциональная зависимость от программы
- проблема воспроизводимости из непреодолимого теоретического предела превращается в вопрос технического искусства.

Субстанциализм и функционализм

Беннет и Хекер [Bennett & Hacker, Neuroscience and Philosophy]:

- «философы сознания» и значительная часть нейрофизиологов **засоряют научный язык** иллюзорными субстанциалистскими терминами
- **неокартезианство**: на место субстанциального дуализма Декарта ставят структурный дуализм тела - мозга, приписывая ментальные предикаты только мозгу.

Беннет и Хекер: когда мы говорим «я знаю» или «он потерял сознание» **мы не подразумеваем мой или его мозг в качестве подлежащего**. Попробуем заменить личные местоимения в этих выражениях на «мой/его мозг» – получим бессмыслицу.

По нашу сторону железного занавеса

- романтический культ НТР
- безудержный оптимизм в отношении перспектив синтетического моделирования основных человеческих способностей
- в сфере массовой культуры: фильм «Его звали Роберт», песня «Сердце из нейлона»
- дискуссия «физиков» и «лириков»
- наивность и почти религиозная бескомпромиссность позиций

По нашу сторону железного занавеса

«Проблема идеального»:

- **Э. В. Ильенков** vs. **Д. И. Дубровский**

Стилистика дискуссии:

- в меньшей степени взаимная логическая оценка аргументов
- в большей – провозглашение ценностных позиций и взаимные обвинения в незнании Гегеля, с одной стороны, и в нарушении целостности категориальной системы диамата, с другой.

По нашу сторону железного занавеса

Ильенков («Об идолах и идеалах»):

- Мыслящее существо должно быть подвижным, умеющим активно действовать в согласии с формой и расположением всех других тел и существ.
- Оно должно активно изменять, переделывать окружающую его естественную среду, строя из нее свое «неорганическое тело».
- Мышление суть функция активной предметной деятельности.
- Мозг – необходимое, но не достаточное условие интеллекта.

По нашу сторону железного занавеса

Дубровский:

- Ильенков должен был бы объявить мимолётные образы или чувства материальными.
- Ильенков должен был бы считать идеальными множественные примеры «представленности» одних явлений в других в живой и неживой природе.

По нашу сторону железного занавеса

Дубровский => «информационный подход»:

(1) информация есть отражение одних материальных систем в других

(2) она не существует вне и помимо материального субстрата, который является также и её «кодом», и по отношению к которому (или которым) она инвариантна

(3) информация в отношении своего субстрата может выполнять функцию управления, что понимается им на основе концепции «информационной причинности»

По нашу сторону железного занавеса

Дубровский:

- «всякое явление сознания есть информация»
- «всякое явление сознания есть функция головного мозга»
- => сознание относится к мозгу как информация к своему материальному носителю («субстрату»).

По нашу сторону железного занавеса

Дубровский и Ильенков **в отношении возможности ИИ**:

- Поверхностный ответ: Ильенков против, Дубровский – за.

Однако:

- концептуально ничто не мешает построить компьютерную модель интеллектуального решения задач как мультиагентную систему
- => реализация ИИ возможна на ильенковских принципах.

По нашу сторону железного занавеса

Предполагаемый ИИ-оптимизм Дубровского тоже может разбиться о его же концепцию мозга как «собственной» кодирующей системы сознания:

- Если сознание возможно только «в материале» мозга, то компьютеры должны разделить второстепенную техническую роль дополнительных кодирующих систем.
- Если же мозг в принципе возможно моделировать в металле, то он лишается своего априори особого положения как привилегированной кодирующей системы сознания.
- Но тогда: если «собственной кодирующей системой» сознания может быть железка, то почему ею не может быть целостность кодирующих систем культуры?

По нашу сторону железного занавеса

Таким образом, с одной стороны, выбор культуры или мозга в качестве «кодирующей системы» сознания не влечёт принятия, соответственно, *скептической* или *оптимистической позиции в отношении ИИ*, а с другой стороны, концептуальное исследование возможности искусственного интеллекта также не даёт преимуществ ни одной из рассматриваемых позиций.

Сравнительный анализ «измов»

Условия интеллекта:

(1) культура (Ильенков, Витгенштейн (?))

↔

(2) естественное или искуственное воплощение вычислительной программы, манипулирующей символами (Деннет, Хофштадтер, Дубровский (?))

↔

(3) субстрат мозга (Сёрл, Дубровский (?))

Сравнительный анализ «измов»

	Субстанциализм	Функционализм
Индивидуализм	(3)	(2)
Холизм	(1)	?

Против (1) и (3): что значит быть функцией субстрата? Сам субстрат есть определённая структура. Тогда мышление – это функция социальной или атомно-молекулярной структуры. Следовательно, деление на субстанциализм и функционализм иллюзорно, можно говорить только о версиях функционализма.

Сравнительный анализ «измов»

Против (2): согласно Беннету и Хекеру, локализация мышления в мозге приводит *неокартезианству*, когда роль субстанций выполняют тело и мозг, и при этом ментальные предикаты приписываются только последнему.Можно сказать, что в случае с потерей сознания употребление «я» – лишь языковая конвенция: на самом деле речь идёт о неполадках в мозге. Но можно ли сказать, что «знать» и «думать» – это тоже предикаты мозга?

Сравнительный анализ «измов»

Позиция (3) тесно связана с **концепцией интенциональности**. Главная работа мозга – производство интенциональных состояний, основа которых – в биологической чувствительности и активности.

Сомнение vs. Боль:

- Боль – состояние не интенциональное: оно в себе содержит свой собственный предмет. А сомнение – всегда «сомнение в том, что...».
- У них разная «внутренняя грамматика».

Сравнительный анализ «измов»

- Цепочка:

(a) факт + фальсифицирующее его обстоятельство → (b) состояние мозга → (c) чувство сомнения

- Согласно Сёрлу, значением высказывания «Я в сомнении» должны быть (b) или (c)
- Что мы подразумеваем, когда говорим «его сомнение» или «он сомневается»? Мы делаем вероятностное умозаключение?

Сравнительный анализ «измов»

- Призовём на помощь ИИ

Мы конструируем **«сомневающуюся» машину**:

- Перцептивный блок
- Управляющая программа
- Нейронная сеть для моделирования (b) и, возможно, (c)
- Желтая и зелёная лампочки в качестве интерфейса

Возражение финансиста:

- Нейронная сеть – слишком дорого, да и не нужно: лампочки может зажигать и управляющая программа.

Сравнительный анализ «измов»

Ваше возражение:

- Нет уверенности, что машина испытает состояние сомнения
- Актёр, изображающий «сомневающееся поведение», сам, скорее всего, не испытывает сомнение своего персонажа
- => Бихевиористский тезис «сомнение есть сомневающееся поведение» оказывается под ударом.

Сравнительный анализ «измов»

A propos:

- **Витгенштейн о «грамматике боли»**: если бы не было «болевого поведения», то, скорее всего, было бы невозможно научить ребёнка правильно использовать слово «боль» [ФИ, п. 257]. Можно ли научить кого-либо сомневаться, если не существует «сомневающегося поведения»?
- **Антибихевиористское возражение**: а можно ли научить кого-либо правильно говорить «я сомневаюсь», если он/она не испытывает состояние сомнения, или мы не знаем, испытывает ли он/она это состояние?

Сравнительный анализ «измов»

Вернёмся к «сомневающейся машине»:

- Если она в ответ на «воздействие сомнительной ситуации» запускает сложнейшую нейронную сеть, имитирующую состояние человеческого мозга...
- Не похоже ли это на самолёт, который, готовясь к взлёту, вдруг начинает хлопать крыльями, подражая голубю.
- Не делаем ли мы излишний шаг на пути к решению задачи?

Сравнительный анализ «измов»

Сторонник Сёрля:

- включение нейронной сети в технологическую цепочку необходимо, если мы хотим моделировать именно сознательное сомнение, а не его автоматическую имитацию
- Но тогда резонен вопрос: если эта *онтологическая основа сомнения* на самом деле не необходима в случае с машиной, то почему мы решили, что она необходима сомневающемуся человеку?
- Но и бихевиористскому тезису мы уже поставили шах аргументом «от актёра»...

Сравнительный анализ «измов»

Правильное концептуальное выражение проблемы:

- там, где один (человек) усомнится, другой будет слепо верить, а третий... тоже слепо верить, но в прямо противоположное и т. д.
- Если же мы построим два или более экземпляра сомневающейся машины, основываясь на простой технологии: рецептор → программа → лампочка – то они, скорее всего, будут солидарны в выражении своих сомнений в сходных обстоятельствах.
- Е/н парадигма: так вот именно мозг и ответственен за индивидуальные различия и свободу! («информационная причинность» у Дубровского)

Сравнительный анализ «измов»

Действия «простого программиста», если ему нужно добиться различного «сомневающегося поведения» от различных машин в сходных обстоятельствах:

- => вставить рандомайзер в управляющую программу.

Тогда:

- Наличие сомнительных обстоятельств
- Наличие двух и более сомневающихся машин без НС, но с рандомайзерами
- Обязательное включение «другой машины» в проблемную ситуацию
- Интерфейс общения и детерминированность к (общей) цели

=> Это ли не модель коммуникативной ситуации у людей?

Сравнительный анализ «измов»

И тогда:

- интенциональные состояния («знание», «сомнение», «вера» и т. п.) – это не состояния мозга или осуществляющейся в нём программы.
- Более того, это и не модусы поведения, чем искушают возможные бихевиористские интерпретации.
- Это ***модальности***, составляющие специфические ***логические структуры*** различных ***коммуникационных ситуаций***.
- От тяжелых раздумий о «мозговом субстрате» и «внутренних состояниях» избавляемся просто на основании бритвы Оккама.

Сравнительный анализ «измов»

	Субстанциализм	**Функционализм**
Индивидуализм	(3)	(2)
Холизм	(1)	Коммуникационный функционализм

Выводы и дальнейшие интриги

Свободными в своей вере и своих сомнениях нас делают вовсе не нейронные сети, а коммуникационные ситуации.

Выводы и дальнейшие интриги

- Если специалисты по ИИ всё еще нуждаются в советах философов, я бы посоветовал им сконцентрироваться на разработке мультиагентных систем.

Выводы и дальнейшие интриги

Возникает некая интересная этическая импликация:

- «права роботов» как возможных носителей ИИ.
- Мы не рассматриваем разрушение робота как преступление против личности
- Но так же было и с рабами. Аристотель: у раба нет своей воли, его душа подобна душе буйвола, т. е. неполноценна.
- Говоря в кантовских терминах: если робот может обладать «чистым разумом», может ли он также обладать «практическим разумом», т. е. «свободно» выбирать между пользой и долгом?
- Оставим этот вопрос для дальнейших исследований.

Bye!

Спасибо за внимание.

Игорь Михайлов

ifmikhailov@iph.ras.ru