Об искусственном интеллекте, его восприятии мира и наделённости сознанием

Введение

В начале был тест, и тест был тестом Тьюринга. Однажды мы задумались о том, что может творение рук наших, и начали задаваться вопросом о том, может ли машина думать. Тьюринг написал о своём знаменитом тесте [Тьюринг 1950], Азимов (хоть и не философ) сформулировал в своих фантастических рассказах три закона робототехники, Сёрль описал мысленный эксперимент с китайской комнатой [Сёрл 1998], Фрэнк Джексон написал о проблеме Мэри и красного цвета [Джексон 1986] – однако происходило это всё относительно давно. Например, в китайской комнате у нас есть некоторый сколь угодно сложный алгоритм, который предоставлен извне – но это алгоритм. Что же существует ныне? Творения рук наших умеют выводить глобальные закономерности обучающей выборки сами – нужно только показать им кусочек данных. Да, это основано на сложной математике, но это не вполне готовый алгоритм.

Теперь машины умеют переводить на китайский почти без ошибок и отличать красный от других цветов по виду. И мы сталкиваемся, кажется, с более интересной проблемой – чтобы определить, каким требованиям должно отвечать сознание искусственного интеллекта, чтобы признать этот искусственный интеллект имеющим сознание.

Итак, какие взгляды на восприятие искусственным интеллектом мира и, соответственно, на наделённость сознанием искусственного интеллекта существуют? Давайте обратимся к истории вопроса.

Тьюринг и его тест

Алан Тьюринг, конечно, не был первым учёным, рассматривающим проблему наделённости искусственного интеллекта сознанием. Однако тест Тьюринга является одной из важных вех развития темы искусственного интеллекта. Тьюринг в своей статье [Тьюринг 1950] сосредотачивает внимание на вопросе «Может ли машина мыслить?». Однако, поскольку термины, в которых задаётся вопрос, необходимо определить. Поэтому Тьюринг говорит о том, что, если мы хотим узнать, действительно ли мыслит данный человек, нам необходимо стать этим человеком – то есть привлекает солипсистский аргумент о том, что мы судим о том, мыслит ли некто, по его действиям, которые кажутся или не кажутся нам осмысленными. Также Тьюринг определяет класс рассматриваемых под словом «машины» объектов – и замечает, что в это множество мы не должны включать людей или клонов людей. Это множество содержит в себе «электронно-вычислительные машины» -- то есть те машины, которые (на своём элементарном уровне) способны выполнить все те же действия, что и человек, выполняющий вычислительные операции (без принятия в рассмотрение объёма бумаги и объёма памяти, а также времени, которое затрачивается на вычисление). То есть все компьютеры и их предки, начиная с вычислительной машины Бэббиджа. Тьюринг переформулирует вопрос в связи с этими двумя уточнениями как «Существуют ли воображаемые цифровые вычислительные машины, которые могли бы хорошо играть в имитацию?». Тьюринг утверждает, что через 50 лет эти машины станут реальностью, и случайный человек за пять минут не сможет определить, говорит ли он с машиной, с вероятностью 70%.

Поскольку Тьюринг много занимался алгоритмикой и математикой, он также рассматривает алгоритмическую сторону того, что это возможно, и упоминает

возможность обучающихся машин. Это достаточно важно для нашего рассмотрения, поэтому давайте обратимся к этому дальше, после описания иных подходов к этой проблеме.

Сёрль и китайская комната

Сёрль [Сёрл 1998] рассматривал проблему уже не в формулировке машины, но в формулировке искусственного интеллекта. Он разделяет ИИ (искусственный интеллект) на сильный и слабый. Согласно слабому, компьютер – мощный инструмент проверки гипотез, сильному – компьютер и есть некоторое сознание, которое обладает набором когнитивных состояний. Какова же суть мысленного эксперимента, который предлагает нам Сёрль? Оставим за скобками этичность представления здесь конкретного языка.

Итак, наш герой эксперимента находится в комнате, у него есть рукопись, текст и вопросы (названий этих трёх объектов наш герой не знает), а также правила ко всем на английском языке, который он понимает, и его просят согласно этим правилам выдавать некоторые китайские символы в ответ на третью рукопись. Согласно Сёрлю, если это происходит с английским языком, он понимает происходящее, если с китайским – нет. Таким образом Сёрль показывает, что, с его точки зрения, то, что претендует на звание «сильного ИИ» не обладает пониманием в том смысле, в котором им обладает человек, даже если внешнему наблюдателю кажется, что это так. Далее Сёрль в пяти частях отвечает на разнообразную критику.

Давайте тоже взглянем на это с критикой (практически солипсисткой критикой). Предположим, что мы создали сильный ИИ, который проявляет инициативу, рассказывает о своих мыслях, имеет желания и так далее. Поверим ли мы ему? Если да, то почему? И почему, если нет? Сёрль отвечает на это и да, и нет – с одной стороны, мы можем, с другой – если мы знаем, что у него есть формальная программа, то мы не должны. Однако, что делать с такими формальными программами, которые являются программами обучения? Интенциональность, как её трактует Сёрль – «свойство определённых ментальных состояний, в силу которого они направлены на объекты и положения дел в мире или в силу которого они суть об этих объектах и положениях дел. Таким образом, полагания, желания и намерения суть интенциональные состояния». Но что такое полагания, желания и намерения и чем они вызваны в отношении человека? Мы, по Сёрлю, мыслящие машины – это так. Цифровой компьютер с подходящей программой, по Сёрлю, не мыслящая машина. Но то, можем ли мы с этим согласиться, зависит от того, что такое «программа» в этом утверждении. К этому мы также придём после литературного обзора.

Функционализм

Здесь мы можем рассмотреть другую точку зрения на вопрос – например, обратимся к Присту [Прист, 2000]. По Присту, нам не очень важно то, в какой именно сложной системе возникает сознание – ментальные состояния обуславливаются некоторыми причинами (тем, что мы имеем на входе) и некоторыми следствиями, которые мы видим на выходе. И тогда основание и конкретная форма этой системы не влияет на то состояние, которое этим вызывается. Почему, как мне кажется, это наиболее подходящее и возможное описание для сознания ИИ, если мы будем создавать его максимально

абстрактным, то есть, не просто решить не очень сложную (или сложную) задачу, а если мы поставим себе цель добиться сознания или приблизиться к нему?

Во-первых, в тех терминах, которые уже были обозначены, и по заветам солипсизма, мы не сможем проверить, действительно ли у него есть сознание или он притворяется. Точно так же, как мы не можем это с точностью установить для человека. Только с конкретной формулировкой и конкретным способом оценивания, в котором учтены все детали и мелочи, мы сможем дать однозначный ответ.

Нужна ли нашему ИИ реальность для этого? Сложный вопрос. С одной стороны, мы модем столкнуться с проблемой Мэри и красного цвета [Джексон 1986] – то есть с тем, что вероятно, если мы будем обучать этот ИИ в ограниченном пространстве, то нам сложно будет рассказать, какая реакция наиболее вероятна при его расширении. Если мы просто расскажем ИИ про весь мир, будет ли он представлять себе его так же, как мы? (Это уже вопрос к лингвистам. Если дать модели очень много текста, как будут выглядеть её представления о мире? В принципе, мы можем попробовать это выяснить, и это интересно. Это не является познанием модели мира человека, но зато является познанием модели мира модели!). С другой стороны и в другом эксперименте, что будет, если мы заставим модель со всеми возможными сенсорами обучаться так, как делает это человеческий ребёнок? И не получим ли мы в результате нечто очень странное и не очень этичное с той точки зрения, что модель выучит, что она человек, или то, что она модель, обладающая правам и возможностями человека?

Возвращаясь к Тьюрингу и Сёрлю, теперь у нас есть программа, которая способна «обучаться». И она будет продолжать «обучаться», и пока мы с ней говорим, и пока она решает дилемму того, что же значат эти карточки на китайском. Но нам необходимо определиться с тем, чего мы хотим достичь. Если мы хотим увидеть у модели разум и сознание в нашем, именно в человеческом понимании, нам придётся создавать машине такие же условия, как у человека. Представим совершенно бесчеловечный и негуманный эксперимент – почти как у Патнема [Патнем, 2002], только интересней. Мы извлекли мозги человека и (каким-то образом) очистили всю память. Или искусственно вырастили мозги, которые не помнят ничего сенсорного. Положили эти мозги в банку, в питательный раствор – им не нужно думать про питание. И придумали механизм делать мозгам неприятно. Итак, мы подключим эту банку к компьютеру и скажем: «Теперь вы будете процессор!». За правильную реакцию награждаем (неважно как), за неправильную – наказываем (тоже неважно как). Вопрос номер один – будет ли это работать процессором. Ну, возможно, будет. А будут ли у этого желания и потребности? Вероятно, всё зависит от задачи. С одной стороны, мы можем использовать это просто как процессор – операционную систему запустить, посчитать что-нибудь. Очень бесчеловечно. Тогда, возможно, даже наша структура, 100% человеческий мозг, не будет обладать разумом. А вот если на другом компьютере запустить другую программу, которая будет эмулировать для наших мозгов окружающий мир и людей – возможно! Может быть, оно будет работать как ничего не знающий о мире, но постепенно узнающий. Таким образом, мозг нашего героя в китайской комнате может и не испытывать вопроса о том, понимает ли он китайский – если он ничего другого в жизни не видел. А может и китайскому научиться, если входной сигнал ассоциирован с чем-то, что уже выучено на предыдущих этапах. Иными словами, если от вас требуют на картинку человека показывать иероглиф человека, и дальше по усложнению, то китайский вы всё-таки выучите.

Заключение и тезисы

Прежде чем сказать, обладает ли что бы то ни было разумом, необходимо определиться с тем, что такое разум. Сёрль говорит о том, что животные разумны, и не отказывает им в разуме. Но если так, до какого уровня пищевой цепи нам необходимо спуститься, чтобы начать отказывать в разуме существам? До инфузории-туфельки, имеющей органоиды, которые позволяют ей реагировать на свет и тень, а также отдёргиваться от соли? До тех червей, у которых появляются нервные клетки? Как мы понимаем, кто обладает интенциональностью, а кто нет? Например, если мы можем некой программой смоделировать все процессы инфузории, и оно будет хищником, как инфузория, и реагировать на свет, как инфузория – получим ли мы что-нибудь?

Конечно, мы имеем некоторые представления о разуме и сознании. Однако нам необходимо определять их формально и однозначно в постановке вопроса, чтобы ответить на него – например, как Тьюринг ответил на переформулированный вопрос.

Литература:

Джексон 1986 – Jackson, Frank (1986). "What Mary Didn't Know". Journal of Philosophy. 83 (5): 291–295

Патнем, 2002 – Патнэм Х. Разум, истина и история[прим. 1] = Reason, Truth and history / Пер. с англ. Т.А. Дмитриева, М.В. Лебедева.. — М.: Праксис, 2002. — 296 с.

Прист, 2000 – Прист, Стивен. Теории сознания / Перевод с английского и предисловие: А. Ф. Грязнов. — Москва: Идея-Пресс, Дом интеллектуальной книги, 2000. — 288 с.

Сёрл 1998 – Дж. Р. Сознание, мозг и программы // Аналитическая философия: становление и развитие. М., 1998.

Тьюринг 1950 – Turing, Alan, «Computing Machinery and Intelligence», Mind LIX (236): 433—460, 1950