Секция «Философия техники»

Критика Бернхардом Иррангом теста Тьюринга: продолжение китайской комнаты Джона Сёрля

Научный руководитель – Яковлев Алексей Александрович

Иванова Елизавета Дмитриевна
Студент (бакалавр)
Российский университет дружбы народов, Факультет физико-математических и естественных наук, Москва, Россия
E-mail: iveeliz.100@gmail.com

Прошло более 70 лет с момента публикации знаменитой статьи Алана Тьюринга «Computing Machinery and Intelligence», в которой автор пытается ответить на вопрос «Может ли машина мыслить?». Но до сих пор тест Тьюринга остается одной из самых влиятельных и спорных тем в области искусственного интеллекта, философии сознания и когнитивной науки. Отношение к тесту Тьюринга складывается разное: есть сторонники и противники. Мы ознакомимся с критикой теста Тьюринга, основанной на аргументах Джона Сёрля и Бернхарда Ирранга. На основе мысленного эксперимента «китайская комната», который Джон Сёрл представил в статье «Minds, Brains, and Programs», проведём реальный эксперимент. Под машиной мы будем понимать среди прочего нейронную сеть.

1. Тест Тьюринга

Для того, чтобы рассматривать критику теста Тьюринга, нам необходимо понять, что он из себя представляет. В начале своей статьи «Computing Machinery and Intelligence» Алан Тьюринг поставил перед собой вопрос: «Может ли машина мыслить?» [5]. Затем, после некоторых размышлений о значении слов «мыслить» и «машина», он предложил ответить на вопрос, может ли машина не хуже человека играть в так называемую «игру в имитацию».

Разберёмся в правилах «игры в имитацию» на основе самой примитивной версии теста. В игре принимают участие: человек, машина и экзаменатор, который тоже является человеком. Экзаменатор находится в отдельной комнате и в письменной форме взаимодействует с машиной и человеком, которые пытаются убедить экзаменатора в том, что они являются людьми. Задача экзаменатора: выяснить, кто является человеком, а кто - машиной. Задача машины: ввести экзаменатора в заблуждение. Если экзаменатору не удаётся уверенно и верно ответить, мы считаем, что машина выиграла игру.

2. Мысленный эксперимент «китайская комната»

Джон Сёрл в статье 1980 года «Minds, Brains, and Programs» привёл аргумент против теста Тьюринга с целью опровергнуть утверждение о том, что машина, наделённая искусственным интеллектом, способна обладать дополнительной интенциональностью помимо той, что мы помещаем в систему в виде формальной программы [4]. Этот аргумент известен в виде мысленного эксперимента «китайская комната».

Джон Сёрл попросил представить, что его заперли в комнате и дали три рукописных текста, которые он до этого не видел, на китайском языке: «рукопись», «рассказ» и «вопросы», а также инструкцию («программу») по сопоставлению формальных символов на английском языке. При этом Сёрл не знает китайский язык и не представляет, как выглядит китайский письменный текст. Но он знает английский язык, поэтому может прочитать инструкцию. Сёрл должен сопоставить наборы формальных символов рукописи и рассказа опираясь только на возможность распознать форму китайских символов. Он отрабатывает навык манипуляции формальными символами на высоком уровне и программисты тоже совершенствуют программу (инструкцию). Затем ему нужно дать ответы на вопросы

с помощью инструкции. Ответы Джона Сёрля получаются неотличимыми от тех, которые дал бы носитель китайского языка. То же самое ему нужно сделать с рассказами и вопросами на английском языке.

Итак, отвечая на английском языке, Джон Сёрл ведёт себя как человек, а при ответах на китайском языке он представляет собой реализацию компьютерной программы. В обоих случаях человек за пределами комнаты будет считать, что в комнате находится человек, хотя во втором случае человек исполняет роль компьютерной программы. Следовательно, тест Тьюринга будет пройден.

Таким образом, Джон Сёрл на примере описанного мысленного эксперимента утверждает, что «тест Тьюринга не является доказательством того, что машина может мыслить», а значит тест Тьюринга не является совершенным способом проверки мыслительных способностей.

3. Не_мысленный эксперимент

Проведём реальный эксперимент на основе «китайской комнаты» Джона Сёрля.

Попросим 10 студентов, которые не знают китайский язык, по очереди пройти в закрытую комнату и предоставим каждому инструкцию по манипуляции китайскими символами на русском языке. Экзаменатор будет передавать через щель вопросы на китайском языке в текстовом виде и ждать, пока студент передаст свой ответ. Инструкция будет составлена таким образом, что после применения всех указанных в ней шагов, студент сможет ответить на полученный вопрос китайскими символами.

Вопросы будут составлены на основе сюжета современного мультфильма «Синий Трактор» для детей дошкольного возраста, с которой может быть знаком либо современный ребёнок до 5 лет, либо члены его семьи. Студент не будет знаком с этой темой, если у него нет брата или сестры, которые смотрят этот мультфильм.

Результаты эксперимента: студенты справились с инструкцией, экзаменатор определил студентов, как детей дошкольного возраста. Значит, студенты смогли убедить экзаменатора в том, что они являются не студентами, а детьми, то есть выступили в роли «машины» и прошли тест Тьюринга. При этом студенты не понимали китайские символы и не давали более развернутые ответы или логичные ответы, содержание которых отличается от ответов из инструкции, а значит не продемонстрировали способность обладать дополнительной интенциональностью, помимо той, что мы предоставили им в виде формальной программы (инструкции).

Таким образом, мысленный эксперимент «китайская комната» проводим и его возможно совершенствовать для повышения точности результатов.

4. Критика китайской комнаты

Идея о развитии понимания китайского языка довольно популярна среди критиков китайской комнаты.

Хасан Чагатай в статье «A Fair Version of the Chinese Room» выдвигает аргумент о том, что человек способен выучить китайский язык в китайской комнате с помощью распознавания образов для возникновения понимания, что игнорирует в своей статье Джон Сёрл. Хасан предлагает «несправедливую версию китайской комнаты», где к описаниям китайских символов будут представлены визуальные данные, с помощью которых человек начнёт понимать китайский язык, а значит будет способен осознанно отвечать точными утверждениями о заданных китайских символах, что противоречит аргументу Сёрля [1].

Пол Маккевитт и Го Каймей в статье «From Chinese rooms to Irish rooms: New words on visions for language» призывают к тому, чтобы пространственные и графические анимированные представления объектов и действий были добавлены в словари и базы знаний для систем с искусственным интеллектом. Они верят, что словари будущего не будут ограничены описаниями символического языка прошлого. Тогда проблема с китайской комнатой

Сёрля исчезнет, поскольку машины будут лучше чувствовать значения слов [3].

5. Бернхард Ирранг: казаться и быть

Рассмотрим аргумент Бернхарда Ирранга против статьи Тьюринга: «Машины могут мыслить, но не так как люди. Точно так же, как самолеты могут летать, но не так как птицы» [2]. Ирранг считает, что вопрос «Может ли машина мыслить?» не имеет смысла, «мышление человека» и «мышление машины» имеют разную природу, поэтому могут рассматриваться только отдельно друг от друга. Человек не может стать машиной, чтобы сравнить понятия «мышления человека» и «мышления машины».

Также Бернхард Ирранг считает, что необходимо различать:

1. Робот - нечто неживое;

2. Человек - зачат от человека, имеет живой мозг и телесность.

Таким образом, дискуссия запуталась в диалектике живого и мёртвого, организма и машины. Насколько мы можем знать сегодня, переход между качествами живого и неживого невозможен. Итак, по мнению Бернхарда Ирранга, утверждение теста Тьюринга о том, что компьютер или робот может действовать как человек, невозможно проверить [2].

6. Вывод

Таким образом, мы рассмотрели некоторые стороны критики теста Тьюринга и можем заключить, что существует критика критики, которая, в свою очередь, не отменяет слабых сторон идей Тьюринга.

Мысленный эксперимент «китайская комната» Джона Сёрля проводим в форме реального эксперимента.

Также можно согласиться с утверждениями Бернхарда Ирранга. Я попыталась ответить себе на вопрос «Может ли человек мыслить как машина?», и поняла, что не могу представить себя на месте машины, поскольку моё мышление и любая машина (даже самообучающаяся нейронная сеть, имитирующая сознание человека) - разные системы.

Таким образом, мы видим, что проведение эксперимента «китайская комната» в реальной жизни иллюстрирует, кроме того, и тезисы Бернхарда Ирранга о несопоставимости «мышления человека» и «мышления машины», представляющих из себя разное.

Источники и литература

1) Cagatay H. A Fair Version of the Chinese Room // Problemos, 2019, vol. 96, p. 121–133.

2) Irrgang, Bernhard E. O. Posthumanes Menschsein?: Künstliche Intelligenz, Cyberspace, Roboter, Cyborgs und Designer-Menschen – Anthropologie des künstlichen Menschen im 21. Jahrhundert. Stuttgart, 2005.

3) McKevitt, P. and Guo, C. From Chinese rooms to Irish rooms: New words on visions for language // Artificial intelligence review, 1996, vol. 10, issue 1-2, p. 49–63.

4) McKevitt, P. and Guo, CM. From Chinese rooms to Irish rooms: New words on visions for language // Artificial intelligence review, 1996, vol. 10, issue 1-2, p. 49–63. 4. Searle, John. R. Minds, brains, and programs // Behavioral and Brain Sciences 3 (3), 1980, p. 417–457.

5) Turing A. M. Computing Machinery and Intelligence // Mind 49, 1950, p. 433–460.