Искусственный интеллект сквозь призму языковых моделей

Елизавета Клыкова (образовательная программа «Фундаментальная и компьютерная лингвистика»)

Аннотация

Начиная со второй половины XX века активно обсуждается возможность создания сильного искусственного интеллекта. Последние разработки в области компьютерной лингвистики (в частности, современные языковые модели) позволяют взглянуть на эту проблему под новым углом. В данной работе предпринимается попытка проанализировать языковые модели с точки зрения присущего им искусственного интеллекта, а также рассматривается вопрос о корректности распространенных аргументов «за» и «против» существования сильного ИИ. Нами показано, что зачастую аргументация противников этой идеи основана на подмене понятий.

Ключевые слова: философия искусственного интеллекта, сильный искусственный интеллект, языковые модели, «китайская комната», тест Тьюринга

1. Введение

Возможность существования искусственного интеллекта (ИИ) в широком смысле - вопрос, который на протяжении многих десятилетий волнует как специалистов в области информационных технологий, так и исследователей в других сферах и простых обывателей. Некоторое время считалось, что искусственный интеллект можно приравнять к человеческому, если их деятельность неотличима с точки зрения человека (в этой связи нужно вспомнить знаменитый тест Тьюринга, который более подробно обсуждается ниже). Появление первых языковых моделей, способных порождать связные осмысленные тексты[1], заставило взглянуть на данную проблему по-новому. Языковая деятельность представляется одной из важнейших особенностей, объединяющих людей и отличающих их от животных. Несмотря на отсутствие общепринятой позиции относительно того, является ли языковая способность врожденной или приобретаемой, говорение на языке признается общей чертой людей, живущих в обществе и не имеющих отклонений, которые мешают усвоению и использованию языка (таких, как афазия и т. д.). Из этого следует закономерный вопрос: должны ли мы считать, что искусственно созданная система, способная к порождению речи, имеет собственный, сравнимый с человеческим интеллект?

[1] Здесь и далее под языковыми моделями мы подразумеваем нейросетевые архитектуры, использующие векторные эмбеддинги (т. е. представления слов в виде совокупности чисел). Существуют разные виды эмбеддингов; например, в более ранних Word2Vec-архитектурах все уникальные слова получали одну фиксированную репрезентацию, тогда как современные трансформеры (BERT, GPT) опираются на контекстуальные эмбеддинги, в которых одно и то же слово может быть представлено по-разному в зависимости от контекста.

В настоящей работе мы рассмотрим существующие взгляды на данную проблему и покажем, в каких аспектах и до какой степени языковые модели можно считать носителями интеллекта в том смысле, в котором он приписывается человеку. Кроме того, мы проанализируем корректность различных аргументов за и против существования сильного ИИ в применении к языковым моделям. Наконец, мы продемонстрируем, что само рассмотрение современных (и более старых) архитектур как потенциальных носителей сильного ИИ содержит ряд теоретических ошибок, которые не должны приниматься за свидетельства против потенциального создания такого типа искусственного интеллекта.

2. Анализ проблемы

Термин «искусственный интеллект» впервые прозвучал в рамках Дартмутского семинара 1956 г., однако немаловажные шаги в становлении этой дисциплины были сделаны несколькими годами ранее. Одним из ее основоположников стал английский математик Алан Тьюринг. В 1950 г. он опубликовал статью «Вычислительные машины и разум», в которой представил эмпирический тест, позволяющий измерить сходство поведения машины и человека. В оригинальной формулировке этот тест представляет собой следующий эксперимент: человек и компьютер по очереди обмениваются письменными репликами с человеком-экспертом (evaluator), который знает, что одним из его собеседников является компьютер[2]. Если эксперт не может с уверенностью указать, кто из двух собеседников на самом деле компьютер, такой компьютер считается прошедшим тест.

Подход Тьюринга активно критикуют за его излишнюю бихевиористичность, т. е. упор на то, каким поведение тестируемого алгоритма представляется со стороны, в противовес тому, что на самом деле происходит внутри этого алгоритма. Джон Сёрль в статье «Сознание, мозг и программы» пишет:

Тест Тьюринга типичен для традиции, будучи откровенно бихевиористским и операционалистским, и я считаю, что если бы разработчики ИИ полностью избавились от бихевиоризма и операционализма, значительная часть путаницы между симуляцией и дублированием была бы устранена (здесь и далее перевод наш. - Е. К.)[3].

Данная цитата содержит важнейший тезис о том, что моделирование (или симуляция) – не то же самое, что дублирование. По утверждению Серля, принципиальное отличие этих двух

[2] *Turing A. M.* Computing Machinery and Intelligence // Mind, vol. LIX (236). 1950. P. 434.

[3] *Searle J. R.* Minds, brains, and programs // Behavioral and Brain Sciences 3(3). 1980. P. 423.

понятий состоит в том, что для первого не требуется понимания в «человеческом» смысле. Перенося аргументацию Серля на современное состояние науки, можно сказать, что компьютерное порождение текста на языке X является моделированием и не предполагает понимания, пусть даже результат неотличим от того, что мог бы произвести носитель этого языка.

В качестве иллюстрации своей позиции Сёрль предлагает мысленный эксперимент, получивший название «китайская комната». В этом эксперименте человек, не знающий китайского языка и не знакомый с его письменностью, получает ряд китайских иероглифов, в ответ на которые пишет другие иероглифы, руководствуясь имеющимся у него сводом правил. Производимые таким образом тексты оказываются неотличимы от результатов языковой деятельности носителя китайского языка, однако создавший их человек по-прежнему не знает китайского и не понимает ни исходного, ни конечного текстов. Если заменить человека компьютерным алгоритмом, то, по мнению Серля, такой алгоритм не будет обладать пониманием (и, шире, мышлением), несмотря на успешное прохождение теста Тьюринга[4].

Важно отметить, что возражения Серля относятся к типу ИИ, который он называет «сильным», т. е. к ИИ, обладающему «когнитивными состояниями» (другими словами, сознанием в человеческом смысле). Сегодня такой искусственный интеллект принято обозначать аббревиатурой AGI (Artificial General Intelligence, общий ИИ, ИИ общего назначения). Существование слабого ИИ (weak или narrow AI), представляющего собой лишь инструмент для решения определенной задачи, на сегодняшний день признается неоспоримым и в настоящей работе не обсуждается.

Итак, рассмотрим принцип функционирования современных языковых моделей в контексте эксперимента Серля. На первый взгляд мы находим почти полное сходство между алгоритмами работы языковой модели и человека в «китайской комнате»: в обоих случаях исполнитель принимает на вход какой-то текст в виде последовательности символов и по определенным правилам конструирует ответ, представляющий собой некоторую новую последовательность. Но если задуматься о том, откуда берутся используемые правила и как они на самом деле применяются, становится очевидным ряд различий. Так, в эксперименте Серля у исполнителя имеется фиксированный, заранее заданный свод правил, по которым порождается текст. В современных языковых моделях такой подход не используется: во-

[4] Там же. С.417–418.

первых, использование строгих правил при порождении текста показывает плохие результаты, а во-вторых (и это положение в некотором смысле является причиной первого), создание исчерпывающего набора таких правил потребовало бы огромного количества ресурсов, как умственных, так и временных, и все равно не привело бы к удовлетворительному результату. Естественный язык не описывается набором разрешений и запретов хотя бы потому, что в любом языке существует вариативность (т. е. в устах разных носителей одна и та же конструкция может звучать по-разному, при этом оставаясь грамматичной и приемлемой с точки зрения других носителей). Распространенным решением этой проблемы в компьютерной лингвистике является обучение языковых моделей на текстах, созданных людьми (носителями языка). Наличие этапа обучения составляет ключевое отличие языковых моделей от исполнителя, помещенного в «китайскую комнату».

Здесь уместно вспомнить утверждение, сформулированное еще графиней Лавлейс в 1843 г.[5] и цитируемое в работах Д. Хартри[6] и А. Тьюринга[7]. Комментируя аналитическую машину Бэббиджа, графиня Лавлейс пишет: «Аналитическая машина не имеет никаких претензий создавать что-либо. Она может делать все, что мы знаем, как приказать ей выполнять»[8]. Д. Хартри соглашается с этим утверждением, приводя в пример генерацию простых чисел при помощи компьютера. Таблица таких чисел, пишет Хартри, легко может быть составлена, если мы научим компьютер выявлять их[9]. Тьюринг предлагает воспринимать этот тезис как утверждение, что компьютер неспособен удивить нас (предоставив нам какую-либо новую информацию). В такой формулировке данный аргумент может быть опровергнут, поскольку понимание какого-либо положения не влечет за собой понимания всех его следствий, и таким образом следствия этого уже известного тезиса могут восприниматься как новая информация[10].

Справедливы ли аргументы Лавлейс и Хартри по отношению к языковым моделям? Эксперименты позволяют утверждать обратное. В разработке нейросетей активно

---

[5] *King A. A., Countess of Lovelace*. Notes by the Translator / L. F. Menabrea. Sketch of the Analytical Engine Invented by Charles Babbage … with Notes by the Translator. London: R. & J. E. Taylor, 1843. P. 691–731.

[6] *Hartree D. R.* Calculating Instruments and Machines. University of Illinois Press, 1949.

[7] *Turing A. M*. Computing Machinery and Intelligence // Mind, vol. LIX (236). 1950. P. 433–460.

[8] *King A. A., Countess of Lovelace*. Notes by the Translator / L. F. Menabrea. Sketch of the Analytical Engine Invented by Charles Babbage … with Notes by the Translator. London: R. & J. E. Taylor, 1843. P. 722.

[9] *Hartree D. R.* Calculating Instruments and Machines. University of Illinois Press, 1949. P.70.

[10] *Turing A. M*. Computing Machinery and Intelligence // Mind, vol. LIX (236). 1950. P. 451.

применяется так называемое обучение без учителя (unsupervised learning), при котором модель получает на вход массив данных и самостоятельно выявляет закономерности, а затем использует их для решения поставленной задачи (например, кластеризации, т. е. разделения объектов на группы по сходству). Применение таких методов в лингвистике долгое время представлялось проблематичным, но уже в 2001 г. предпринимаются относительно успешные попытки анализа синтаксической структуры с помощью обучения без учителя[11]. Сегодня этот тип обучения используется в самых различных структурах от векторных моделей типа Word2Vec, впервые представленных в 2013 г.[12], до глубоких нейронных сетей, порождающих текст (на момент написания данной работы наиболее продвинутой из моделей этого типа является GPT-3 для английского языка[13]). Интересно, что такие модели способны анализировать данные на принципиально иных уровнях, чем это делает человек, и таким образом обнаруживать и усваивать свойства, недоступные разработчикам. Более того, в некоторых случаях алгоритмы срабатывают лучше, чем ожидает их создатель, находя закономерности там, где их не ожидается. Таким образом, аргументы Лавлейс и Хартри оказываются неверными в отношении современных языковых моделей.

Итак, языковые модели умеют обучаться и формулировать ранее неизвестные закономерности (говоря в терминах А. Тьюринга, они могут удивлять создателей). В процессе обучения и работы модели неизбежно ошибаются, - это еще одно свойство, которое роднит их с человеческим сознанием. Примечательно, что неспособность машин ошибаться активно обсуждалась как один из аргументов против существования сильного ИИ. В статье А. Тьюринга «Вычислительные машины и разум» говорится, что компьютер можно научить «ошибаться», т. е. иногда выдавать неверные ответы, чтобы ввести эксперта в заблуждение и внушить ему, что компьютер на самом деле человек[14]. При этом невозможность ошибок в работе компьютера как абстрактного алгоритма не оспаривается:

---

[11] *Clark A. S.* Unsupervised Language Acquisition: Theory and Practice. PhD dissertation. University of Sussex, 2001.

[12] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013. P. 1–12.

[13] *Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D.* Language Models are Few-Shot Learners // Neural Information Processing Systems (33), 2020. P. 1877–1901.

[14] *Turing A. M.* Computing Machinery and Intelligence // Mind, vol. LIX (236). 1950. P. 448.

Эти абстрактные машины – математические фикции, а не физические объекты. По определению они неспособны совершать ошибки функционирования. В этом смысле мы действительно можем сказать, что «машины никогда не могут ошибаться»[15].

В случае с языковыми моделями ситуация обстоит принципиально иначе: цель разработчиков – добиться как можно большей точности, а не научить алгоритм симулировать ошибки, однако ни одна из существующих архитектур не дает идеального качества. Более того, на сегодняшний день нет моделей, способных ввести в заблуждение хотя бы 30% судей (такой порог был задан самим Тьюрингом). Казалось, что в 2014 г. чат-бот Eugene Goostman[16] перешагнул этот порог, «обманув» одного из трех экспертов, но в связи с многочисленными методологическими нарушениями, допущенными при проведении эксперимента, тест Тьюринга нельзя считать пройденным[17][18]. Сегодня наиболее совершенным представляется чат-бот Mitsuku (сокращенно Kuki)[19], побеждавший в ежегодном конкурсе «AI Loebner» пять раз (в 2013 и 2016–2019 гг.)[20]. Однако приз размером в $25000, предназначенный первому алгоритму, который сможет ввести в заблуждение хотя бы половину членов жюри, еще не был вручен.

Несомненно, процессы усвоения языка человеком и моделью различаются. Для человека обучение языку неразрывно связано с коммуникацией, в ходе которой он запоминает не только то, какие слова и в каком порядке составляют грамматичные предложения, но и то, какие последствия (в широком смысле) имеют те или иные высказывания. Языковая модель лишена второго аспекта; единственная реакция, которая доступна ей во время обучения, это штраф за неверные ответы, корректировка гиперпараметров и т. д. Возвращаясь к аргументу Серля о непонимании компьютером языка, на котором он порождает тексты, можно сказать, что модель действительно не обучается пониманию: получая штрафы, она «запоминает» ошибки

---

15 Там же. С. 449.

16 https://web.archive.org/web/20140703103434/http://www.princetonai.com/bot/ (на момент обращения, 12.12.2021, сайт недоступен).

17 *Masnick M.* No, a 'Supercomputer' Did NOT Pass the Turing Test for the First Time and Everyone Should Know Better. TechDirt, Jun. 3, 2014. URL: https://www.techdirt.com/articles/20140609/07284327524/no-supercomputer-did-not-pass-turing-test-first-time-everyone-should-know-better.shtml Просмотрено 12.12,2021.

18 *Moesser M.* Chat bots and the Turing test. Oxford Protein Informatics Group, Sep. 16, 2019. URL: https://www.blopig.com/blog/2019/09/chat-bots-and-the-turing-test/ Просмотрено 12.12.2021.

19 Информация о боте доступна по ссылке https://www.kuki.ai/research, «пообщаться» с ним можно по ссылке https://www.kuki.ai/.

20 Важно отметить, что с 2019 г. соревнование больше не проводилось.

и учится их не повторять, но причины ей неизвестны. Попробуем, однако, сравнить этот процесс с усвоением языка человеком. Изучая иностранный язык, мы часто слышим от преподавателей, что то или иное правило или конструкцию нужно просто запомнить, поскольку понять их невозможно. Можно ли сказать, что мы при этом не обладаем пониманием изучаемой части языка? Следуя аргументации Серля, мы должны считать, что человек, говорящий на языке, делает это осмысленно; но, как мы показали, человеку не всегда доступно понимание того, почему язык функционирует так, а не иначе. Это верно не только для иностранного, но и для родного языка.

Однако на приведенное рассуждение легко возразить, что важно не только и не столько понимание того, как мы используем язык, сколько того, почему и зачем мы это делаем. Отсутствие у компьютерной программы интенциональности и каузальных свойств, имеющихся у человеческого сознания, – основной аргумент Серля против существования сильного ИИ. С этим утверждением трудно поспорить; можно говорить, например, что интенциональность человека преувеличена и что многие наши действия являются лишь закономерной реакцией на внешние стимулы, но факт остается фактом: выбирая, например, между игнорированием оскорбления и ответом на него, мы принимает волевое решение, чего не способна сделать языковая модель, если только мы не встроили в нее такую возможность. Однако именно в последней фразе кроется суть нашего следующего аргумента: невозможно сравнивать языковые модели (и в целом искусственный интеллект) с человеческим сознанием, используя ту конфигурацию, которая распространена на сегодняшний день. Разрабатывая тот или иной алгоритм, мы закладываем в него набор функций, которые он должен выполнять, и выйти за рамки этого набора программа неспособна - мы просто не даем ей такой возможности.

В своей критике «Ответа от нескольких обиталищ» Сёрль не отрицает перспективу создания в будущем искусственного интеллекта, не ограниченного возможностями современных ему компьютеров, однако пишет, что такой подход «тривиализует замысел сильного ИИ, переопределяя его как все, что искусственно продуцирует и объясняет познание»[21]. Такая постановка проблемы сводится к вопросу о том, что подразумевается под сильным ИИ, а не о том, возможно ли его существование в принципе. Ограничиваясь

[21] « [This approach] trivializes the project of strong AI by redefining it as whatever artificially produces and explains cognition.» [Searle J. R. Minds, brains, and programs // Behavioral and Brain Sciences 3(3). 1980. P. 422.]

исходным определением, мы, пожалуй, должны признать, что сильный искусственный интеллект не может существовать; но он не может существовать потому, что мы задумали его таким. Развивая пример Серля с фотосинтезом, можно представить рассматриваемую проблему через следующую аналогию: ученые разработали вещество, потенциально способное к фотосинтезу, и сравнивают его с хлорофиллом. При этом проводится наблюдение за колбой с полученной субстанцией, находящейся в подземной лаборатории, и за цветком, растущим на улице под открытым небом. В данном примере колба относится к цветку так, как нейронная сеть относится к человеческому организму, и это отношение демонстрирует различия в сложности сравниваемых структур. Расположение колбы и цветка под землей и на улице соответственно отражает различия в возможностях развития (обучения) искусственного интеллекта и человека.

Идея, созвучная со сформулированной нами выше, обсуждается Р. В. Душкиным. Автор указывает на методологические ошибки, присутствующие в эксперименте с «китайской комнатой»:

[В] этом мысленном эксперименте была попытка обосновать невозможность построения сильного ИИ, в то время как в пример приводилась схема слабого ИИ-агента, основанного на вычислительном подходе в рамках архитектуры фон Неймана. Это выглядит примерно так же, как если бы в отношении человека разумного был бы задан вопрос «понимает ли какой-либо конкретный нейрон […] то, что этот человек в заданный момент времени читает?»[22].

В работе предлагается гибридная архитектура искусственного интеллекта, которая, по мнению Р. В. Душкина, может лечь в основу действительного сильного ИИ. При этом автор выступает за отказ от применения эксперимента Серля к современным архитектурам ИИ-агентов.

3. Заключение

В данной работе мы проанализировали некоторые из наиболее известных взглядов на проблему сильного ИИ, рассмотрев их в контексте современных языковых моделей. Наши выводы можно распространить и на прочие алгоритмы, задействующие нейросетевые архитектуры (потенциально – и на другие виды слабого ИИ, но это обсуждение остается за рамками нашей работы). Мы показали, что языковые модели имеют больше общего с

[22] *Душкин Р. В.* Критика «Китайской комнаты» Дж. Сёрла с позиции гибридной модели построения искусственных когнитивных агентов // Сибирский философский журнал 18(2), 2020. С. 32.

человеческим сознанием и стоят ближе к сильному ИИ, чем алгоритмы, существовавшие в XX веке.

Однако мы также продемонстрировали, что подобные алгоритмы по определению не могут удовлетворять всем требованиям, предъявляемым к сильному искусственному интеллекту. Из этого следует, что использование «китайской комнаты» Дж. Серля как аргумента о невозможности разработки сильного ИИ некорректно.

Изначальная проблема воспроизводимости человеческого сознания (точнее, сознания в человеческом смысле) техническими средствами остается открытой, как и вопрос о том, какая именно часть сознания (и насколько большая) должна присутствовать у машины, чтобы ее можно было признать обладающей интеллектом[23]. Эти аспекты требуют дополнительного обсуждения.

Библиография

Душкин Р. В. Критика «Китайской комнаты» Дж. Сёрла с позиции гибридной модели построения искусственных когни-тивных агентов // Сибирский философский журнал 18(2), 2020. С. 30–47.

Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. Language Models are Few-Shot Learners // Neural Information Processing Systems (33), 2020. P. 1877–1901.

Clark A. S. Unsupervised Language Acquisition: Theory and Practice. PhD dissertation. University of Sussex, 2001.

Hartree D. R. Calculating Instruments and Machines. Uni-versity of Illinois Press, 1949.

King A. A., Countess of Lovelace. Notes by the Transla-tor / L. F. Menabrea. Sketch of the Analytical Engine Invented by Charles Babbage … with Notes by the Translator. London: R. & J. E. Taylor, 1843. P. 691–731.

Masnick M. No, a 'Supercomputer' Did NOT Pass the Turing Test for the First Time and Everyone Should Know Better. TechDirt, Jun. 3, 2014. URL:

[23] Похоже, что ответ на этот вопрос зависит в том числе от уровня обобщения: например, если мы считаем, что сознание каждого человека уникально, то повторить его, безусловно, представляется невозможным.

https://www.techdirt.com/articles/20140609/07284327524/no-supercomputer-did-not-pass-turing-test-first-time-everyone-should-know-better.shtml Просмотрено 12.12.2021.

Mikolov T., Chen K., Corrado G., Dean J. Efficient Estima-tion of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013. P. 1–12.

Moesser M. Chat bots and the Turing test. Oxford Protein In-formatics Group, Sep. 16, 2019. URL: https://www.blopig.com/blog/2019/09/chat-bots-and-the-turing-test/ Просмотрено 12.12.2021.

Searle J. R. Minds, brains, and programs // Behavioral and Brain Sciences 3(3). 1980. P. 417–424.

Turing A. M. Computing Machinery and Intelligence // Mind, vol. LIX (236). 1950. P. 433–460