Министерство образования и науки Российской Федерации
Федеральное агентство по образованию
Федеральное государственное образовательное учреждение высшего профессионального образования «Санкт-Петербургский государственный университет»
Математико-механический факультет
Кафедра математической кибернетики

Думающие машины

Сарапулов Георгий Владимирович

Заведующий кафедрой:
д. ф.-м. н., профессор А. Л. Фрадков
Научный руководитель:
к. ф.-м. н., доцент М. С. Ананьевский

Санкт-Петербург
2018

Содержание

Введение

В реферате предпринята попытка сделать обзор идей, возникших в обсуждении вопроса о способности машин мыслить. Наиболее активная дискуссия развернулась вокруг двух известных мысленных экспериментов. С одной стороны - это имитационная игра Алана Тьюринга, предложившего способ выявить способность компьютера думать и верившего, что в скором будущем машины смогут успешно проходить предложенный им тест на способность к мышлению. С другой стороны - это Китайская Комната Джона Сёрл, занимавшего скептическую позицию по отношению к возможности создания так называемого сильного искусственного интеллекта. Эти мысленные эксперименты вызвали большой интерес исследователей и философов, и несмотря на то, что имитационной игре Тьюринга уже почти семьдесят лет, а Китайской комнате - почти полвека, споры ведутся до сих пор. Так, имитационная игра Тьюринга побудила огромное число исследовательских групп предпринять попытки построить диалоговые системы, способные убедить человека в том, что он беседует с человеком, не с машиной; следует отметить, что пока безуспешно, вопреки прогнозам самого Тьюринга, однако эти попытки позволили добиться больших достижений в обработке естественного языка и привлекли внимание ученых к проблеме понимания механизмов, которые позволяют человеку вести диалог с собеседником и которых не хватает машине, чтобы делать это так же естественно. В свою очередь эффект, вызванный Китайской комнатой и теми философскими проблемами, которые он обозначил, был так значителен, что Патрик Хайес, ныне возглавляющий Институт Человеческого и Машинного Познания во Флориде, предложил переопределить когнитивистику как «продолжающуюся исследовательскую программу по опровержению аргумента Сёрла»[1].

В первой части излагаются идеи, описанные Тьюрингом в статье «Вычислительные машины и разум» и другими исследователями и философами в ответ на приведенные аргументы. В ней описан мысленный эксперимент, имитационная игра, которая по мнению Тьюринга позволит ответить на вопрос «Могут ли машины думать?», а также возможные возражения против этой способности, на которые Тьюринг указывал в статье. В противопоставление оптимистичному взгляду Тьюринга на способность машин к

[1] [1] Harnad S. What's wrong and right about searle's chinese room argument? // [Book Chapter] (in Press) / Ed. by Michael A. Bishop, John M. Preston. — Oxford University Press, 2001.

мышлению, вторая часть приводит аргумент Сёрла, утверждавшего о невозможности возникновения понимания из формальной обработки символов, которую осуществляет машина. Подобно Тьюрингу, Сёрл детально обсуждает возможные аргументы против собственных доводов, с той лишь разницей, что позиция Сёрла в вопросе противоположна. В ходе описания ответов на агрумент приводятся также отсылки к более широким философским проблемам, выходящим за рамки искусственного интеллекта.

Глава 1

Имитационная игра Алана Тьюринга

В статье [2] Тьюринг отмечает, что ответ на вопрос «Могут ли машины думать» необходимо было бы начать с определения понятия «машина» и «думать». Вместо этого, чтобы избежать неоднозначности толкований, связанных с повседневным пониманием этих слов, он заменяет исходный вопрос мысленным экспериментом, который он сам назвал «Имитационной игрой» («The Imitation Game»), впоследствии ставшим известным как «тест Тьюринга». Суть первоначальной игры заключалась в том, что один из участников игры задает вопросы двум другим участникам (мужчине и женщине, находящимся в отдельной комнате) чтобы выяснить, кто из них двоих - мужчина, а кто - женщина. Один из опрашиваемых (A) старается обмануть опрашивающего, второй (B), наоборот, старается ему помочь. В рамках мысленного эксперимента Тьюринг заменяет участника A на машину и задает вопрос: будет ли опрашивающий ошибаться в новой игре между человеком и машиной так же часто, как он ошибался бы в игре между мужчиной и женщиной?

Иногда считается, что приведенный Тьюрингом мысленный эксперимент в других формах встречался и ранее. Например, Gunderson [3] приводит отрывок из трактата Декарта «Рассуждение о методе», где Декарт отмечает, что хотя машины могут справляться с рядом задач так же успешно, как человек, некоторые виды деятельности представляют для него непреодолимые трудности, в которых машина неибежно обнаружит, что действует исходя не из понимания, а из расположения своих составляющих, то есть работы своих механизмов. Из дальнейших рассуждений Декарта можно заключить, что, по его мнению, ни одна машина никогда не сможет научиться давать осмысленные ответы на вопросы, поскольку это требует наличия огромного количества внутренних составляющих («органов») для определения своих действий в различных обстоятельствах. Действительно, ход мыслей Декарта удивительно схож с рассуждениями Тьюринга в той части, где Тьюринг касается вопроса количества дискретных состояний машины, необходимых для успешной имитационной игры. Как будет видно далее, в рассуждениях Декарта также прослеживается и идея, положенная в основу Китайской комнаты: машина действует не понимая, а лишь из расположения органов (у Сёрла - лишь осуществляя формальные операции). Тем не менее, по всей видимо-

сти, для Декарта предложенный Тьюрингом тест был бы убедительным свидетельством способности машины думать.

Сам Тьюринг считал, что с уже спустя полвека будет возможно запрограммировать компьютер для игры в имитацию на таком уровне, что средний опрашивающий будет верно идентифицировать собеседников в пятиминутном диалоге в среднем не более чем в 70% случаев. Даже если мозг человека является ялвяется машиной с непрерывными состояниями, Тьюринг полагал, что машина с дискретными состоянями способна имитировать работу нашего мозга достаточно хорошо, чтобы быть успешной в игре. Против такой точки зрения Тьюринг приводит возможные возражения.

Согласно теологическому возражению, мышление порождено бессмертной душой, данной Богом каждому человеку, но не животному или машине, в связи с чем животные и машины мыслить не способны. Этот аргумент согласуется с позицией дуалистов, считающих мышление функцией нематериальной, существующей отдельно от тела сущности. Согласно этой точке зрения, материальное тело не является носителем мысли, и поэтому машины ничем не отличаются от других тел в своей неспособности к мышлению. Указывая на спекулятивный характер рассуждений, Тьюринг указывает, что приведенное выше возражение влечет существенное ограничение власти Всевышнего, который должен быть способен наделить душой как слона (наделив его более совершенным мозгом для обслуживания души), так и машину. При этом создание такой машины узурпировало бы исключительную власть бога на создание душ не в большей степени, чем рождение детей, так как в обоих случаях люди выступают как инструменты его воли, предоставляющие материальное тело для создаваемых им душ.

Образно названное «Головой в песке» возражение относится к опасениям, связанных с возможными ужасными последствиями появления мышления у машин. Это возражение скорее отражает не логические препятствия для возможности машин к мышлению, а различные связанные с ним страхи, поэтому оно также не заслужило особого внимания.

Математические возражения опираются на ряд результатов в математической логике, в частности, теорему Гёделя о неполноте, которые показывают ограниченность возможностей дискретных машин. Суть результатов в том, что внутри достаточно сильной формальной системы существует класс истинных утверждений, которые не могут быть доказаны внутри системы. Тьюринг понимал, что эти результаты влияют на предло-

женный им тест и указывает на тип вопросов, на которые цифровой компьютер с бесконечной памятью неспособен дать ответ: «Представьте себе машину [описание машины]. Сможет ли она ответить 'Да' на любой вопрос?». Математический результат показывает, что если описание машины в какой-то степени схоже с описанием опрашиваемой машины, ответ будет неверным либо его вообще не последует. Однако остается вопрос, является ли свобода от таких ограничений необходимой для способности мыслить, как неясно и то, свободны ли люди от этих ограничений. Кроме того, для составления подобных вопросов необходима детальная информация о внутреннем устройстве машины, и в таком случае речь идет уже не о среднем опрашивающем, который подразумевался в имитационной игре, а о хорошо подготовленном специалисте.

Другая группа возражений связана с самосознанием машины: согласно этой линии аргументации, мы можем признать машину мыслящей, действует исходя из собственных чувств и осознает собственные действия. Такая точка зрения близка к солипсизму, поэтому вокруг нее сложно выстроить дискуссию, ведь у нас нет надежных критериев, чтобы определить, осознают ли себя другие люди. Внешние проявления эмоций, как и прочие человеческие способности, напрямую не связанные с разумностью (быть добрым, дружелюбным, иметь чувство юмора и т. п.) могут не быть свойственными другим разумным существам, поэтому требовать их наличия было бы проявлением шовинизма. К тому же, нет достаточных оснований считать, что соответствующим образом запрограммипрованных компьютер не способен на эти вещи.

Одно из самых популярных возражений берет начало из заметки Ады Лавлейс по поводу аналитической машины Бэббиджа: она не претендует на то, чтобы создавать что-то новое, но следует инструкциям и делает то, что мы способны объяснить ей, как делать. В качестве контраргумента Тьюринг задается вопросом, способен ли человек делать действительно что-то новое, и указывает, что люди также подвержены ограничениям, исходящим из биологии и генетического наследования. Этот контраргумент был воспринят критически. Например, Bringsjord [1] люди производят новые предложения на естественном языке практически в каждом диалоге, в то время как способность машин к этому остается под вопросом.

Другое возражение связано с возможностью, которую признавал сам Тьюринг, что мозг является машиной с непрерывными состояниями. Тьюринг утверждал, что

[1] [4]

даже если это так, то такая машина может быть имитирована машиной с дискретными состояниями с незначительными ошибками. Однако в этом случае, если мышление в действительности свойственно только машинам с непрерывными состояниями, то тест Тьюринга будет вводить в заблуждение, и мы будем иметь дело не с мышлением, но с его имитацией. В то же время, как справедливо отмечает Block [2], в нашей концепции разума нет ничего, что говорит о невозможности существования разумных существ с квантованными сенсорными устройствами или цифровыми рабочими частями.

Предложенный Тьюрингом тест впоследствии подвергался критике. С одной стороны, ряд исследователей считал его слишком сложным: например French [3] писал, что тест бесполезен, поскольку есть вопросы, раскрывающие низкоуровневую структуру сознания, поэтому для успешного прохождения теста пришлось бы ее реализовать, что практически невозможно. Кроме того, могут быть отдельные особенности человеческого мышления, которые очень сложно симулировать, но которые при этом не являются обязательными для мышления. Согласно другой линии возражений, тест Тьюринга является слишком узконаправленным. Как отмечает Gunderson [3], успех в тесте может быть обусловлен причинами иными, чем наличие мышления, и является всего лишь одним примером из того спектра способностей, которыми обладают разумные существа. Есть и исследователи, которые, напротив, считают тест слишком простым. Harnad[4], например, считал более подходящей целью для исследований прохождение Полного Теста Тьюринга (Total Turing Test), подразумевающего, что система должна давать ответы на любые входные данные, не только лингвистические.

[2] [5]
[3] [6]
[4] [7]

Глава 2

Китайская комната Джона Сёрла

Спустя тридцать лет после статьи Тьюринга Джон Сёрл описал другой мысленный эксперимент, который часто упоминается в связи с имитационной игрой и который вывел обсуждение вопросов, связанных с мышлением и искусственным интеллектом на новый уровень. Сёрль ставит под сомнение, что компьютер, запрограммированный на составление ответов, буквально обладает когнитивными состояниями.

[8]

Замечание 1. α

2

Заключение

Литература

1. Harnad S. What's wrong and right about searle's chinese room argument? // [Book Chapter] (in Press) / Ed. by Michael A. Bishop, John M. Preston. — Oxford University Press, 2001.

2. Turing A. M. Computing machinery and intelligence // Mind. — 1950. — Vol. 59, no. 236. — P. 433–460. — URL: http://www.jstor.org/stable/2251299.

3. Gunderson K. Descartes, la mettrie, language, and machines // Philosophy. — 1964. — Vol. 39, no. 149. — P. 193–222.

4. Bringsjord S., Bello P., Ferrucci D. A. Creativity, the turing test, and the (better) lovelace test // Minds and Machines. — 2001. — Vol. 11, no. 1. — P. 3–27.

5. Block N. Psychologism and behaviorism // Philosophical Review. — 1981. — Vol. 90, no. 1. — P. 5–43.

6. French R. M. Subcognition and the limits of the turing test // Mind. — 1990. — Vol. 99, no. 393. — P. 53–66.

7. Harnad S. Minds, machines and searle // Journal of Experimental and Theoretical Artificial Intelligence. — 1989. — Vol. 1, no. 4. — P. 5–25.

8. Searle J. R. Minds, brains, and programs // Behavioral and Brain Sciences. — 1980. — Vol. 3. — P. 417–424.