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 3

Глава 1. Имитационная игра Алана Тьюринга 5

Глава 2. Китайская комната Джона Сёрла 9

Заключение 12

Литература 13

Введение

Исследования ученых в области искусственного интеллекта позволили создавать программы, способные выполнять действия, которые, по нашему опыту, требуют наличия мышления. Спустя двадцать лет после победы Deep Blue над действующим чемпионом мира по шахматам Гарри Каспаровым система AlphaGo сумела уверенно обыграть одного из сильнейших игроков в Го Ли Седоля - в игру, умение играть в которую, как считается, требует наличия не только стратегического мышления, но и интуиции. В свете этих достижений часто ставится вопрос о способностях компьютеров к мышлению и их возможностях превзойти человека во всех сферах мыслительной деятельности.

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 Сёрля, 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 исследо-

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

[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.

Глава 1

Имитационная игра Алана Тьюринга

В статье 1950-го года, опубликованной в философском журнале Mind[1], 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, Гундерсон [2] 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

[1] [2] Turing A. M. Computing machinery and intelligence // Mind. — 1950. — Vol. 59, no. 236. — P. 433–460. — URL: http://www.jstor.org/stable/2251299.

[2] [3] Gunderson K. Descartes, la mettrie, language, and machines // Philosophy. — 1964. — Vol. 39, no. 149. — P. 193–222.

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 Сёрля - лишь осуществляя формальные операции). Тем не менее, по всей видимости, 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. Понимая 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 ло-

гике, 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 предложенный 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, если она 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, Брингсьйорд[3] отмечает, что люди производят новые

[3] [4] Bringsjord S., Bello P., Ferrucci D. A. Creativity, the turing test, and the (better) lovelace test

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 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 Блок[4], 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 Френч[5] писал, что тест бесполезен, поскольку машине в ходе теста можно задавать 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 Гундерсон[6], 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. Харнад[7], 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.

// Minds and Machines. — 2001. — Vol. 11, no. 1. — P. 3–27.

[4] [5] Block N. Psychologism and behaviorism // Philosophical Review. — 1981. — Vol. 90, no. 1. — P. 5–43.

[5] [6] French R. M. Subcognition and the limits of the turing test // Mind. — 1990. — Vol. 99, no. 393. — P. 53–66.

[6] [3] Gunderson K. Descartes, la mettrie, language, and machines // Philosophy. — 1964. — Vol. 39, no. 149. — P. 193–222.

[7] [7] Harnad S. Minds, Machines, and Searle // Journal of Experimental and Theoretical Artificial Intelligence . — 1989. — Vol. 1, no. 4. — P. 5-25.

Глава 2

Китайская комната Джона Сёрла

Спустя тридцать лет после статьи Тьюринга Джон Сёрль описал другой мысленный эксперимент [1], который часто упоминается в связи с имитационной игрой и который вывел обсуждение вопросов, связанных с мышлением и искусственным интеллектом, на новый уровень. В рамках экспреимента Сёрль представляет, что он заперт в комнате с большим корпусом текста на китайском языке (который он не понимает и даже не способен отличить китайские иероглифы от любых других символов) и набором инструкций на английском языке (которые он понимает), описывающих, какие манипуляции над символами необходимо произвести, чтобы в ответ на вновь введенную последовательность символов вернуть сгенерировать другую последовательность. Со временем он настолько хорошо научится следовать инструкциям, что внешние наблюдатели могут прийти к ошибочному выводу, что находящийся в комнате человек, выдающий очень убедительные ответы на задаваемые ему вопросы, понимает китайский язык. Этим экспериментом Сёрль ставит под сомнение адекватность теста Тьюринга: согласно его выводу, компьютер применяет синтаксические правила для манипуляции строками, но не имеет представления о семантике, то есть не обладает пониманием смысла, не способен соотнести значение со словами.

В широкой трактовке аргумент Сёрля говорит о том, что человеческое сознание не является вычислительной системой, подобной компьютерам, но возникло из биологических процессов, которые компьютер способен лишь имитировать. Тем самым он отрицает функционалистский подход к пониманию сознания, в рамках которого когнитивные состояния определяются их причинно-следственными ролями, а не материалом, который выполняет соответствующие функции. В более узком смысле аргумент направлен против концепции «сильного искусственного интеллекта», согласно которой формальные вычисления над символами способны породить мышление, и тем самым компьютеры действительно способны понимать естественный язык и делать осознанные ходы в шахматах. Альтернативная концепция «слабого искуственного интеллекта» не утверждает, что компьютеры способны к мышлению, оставляя им роль полезных

[1] [8] Searle J. R. Minds, brains, and programs // Behavioral and Brain Sciences. — 1980. — Vol. 3. — P. 417–424.

инструментов.

Мы ограничимся рассмотрением критики агрумента в узком смысле. Обычно возражения следуют одной из трех логических линий.

Согласно первой, находящийся в комнате действительно не понимает китайский язык, однако работа системы создает некую сущность, способное к пониманию.Человек в комнате является частью системы вместе с базой текста, инструкциями и результатами промежуточных операций, и хотя он сам не понимает китайский язык, система в целом понимает, либо понимание реализуется в виртуальном сознании. Сёрль отвечает на это тем, в принципе человек может заключить в себе все компоненты системы, запомнив все инструкции и делая все вычисления в уме. Харнад[2] в этой связи также отмечает, что если речь идет о понимании китайского языка самой программой, выполняющей инструкции, то аргумент Сёрля выглядит верным. Более убедительная аргументация связана с тем, что компьютер, исполняющий программу, не эквивалентен самой программе: он является сложной электронной системой с внутренними причинно-следственными связями. И хотя мы можем интерпретировать его физические состояния как синтаксические нули и единицы, в реальности синтаксис является производной от электронной природы, что существенно отличает компьютер от абстрактной формальной системы. Деннетт[3] в этой связи считает, что программирование в действительности способно дать сознание. Он также допускает, что немаловажное значение может иметь скорость выполнения операций, так что оператор китайской комнаты не обладает пониманием, в то время как компьютер в подобной ситуации мог бы понимать китайский язык.

Ученые, придерживающиеся второй линии рассуждения, согласны с Сёрлем в том, что сам процесс выполнения программы не создает понимания, но если систему, исполняющую программу, поместить в тело робота, способного обучаться через наблюдение и взаимодействие с окружающим миром, или в детальную симуляцию мозга, то такая расширенная система будет способна к пониманию. Тем самым признается, что синтаксис и внутренние взаимосвязи недостаточны для получения семантики, в то время как наличие подходящих связей с внешним миром может дать смысловое содержание

[2] [7] Harnad S. Minds, Machines, and Searle // Journal of Experimental and Theoretical Artificial Intelligence . — 1989. — Vol. 1, no. 4. — P. 5-25.

[3] [9] Dennett D. C. Toward a cognitive theory of consciousness // Minnesota Studies in the Philosophy of Science. — 1978. — Vol. 9.

внутренним символьным представлениям. Сёрль возражает, что всевозможные сенсоры дают лишь еще один синтаксический канал ввода. Один из убедительных ответов на аргумент дал Рей[4]. Защищая позиции функционализма, он утверждает, что изолированная система, описанная Сёрлем, не является функционально эквивалентной человеку, говорящему на китайском языке и взаимодействующему с окружающим миром. Он критикует Сёрля за то, что в его рассуждениях функционализм сведен к описанию объекта по принципу черного ящика, что в большей степени свойственно бихейвиоризму, в то время как в рамках функционализма важно внутреннее устройство. В качестве контрпримера Рей описывает систему, у которой есть восприятие, которая способна к дедуктивным и индуктивным выводам, руководствуется целями и представлениями в процессе принятия решений и способна обрабатывать естественный язык, преобразуя его к нативным для себя представлениям и обратно. Объяснение поведения такой системы потребовало бы тех же средств, что и в случае говорящего на китайском языке человека.

Третья линия рассуждения придерживается того, что человек в китайской комнате может понимать китайский, либо описанный сценарий невозможен. Аргументация основывается на том, что интуиция в подобном случае может вводить в заблуждение. Так Пол и Патриция Чёрчланд [5] придерживаются взгляда, что мозг осуществляет векторные операции, не оперируя символами согласно каким-либо структурным правилам, и интуиция подводит, когда мы имеем дело с такими сложными системами при переходе от части к целому. Сюда же можно отнести уже упомянутые выше аргументы по поводу скорости выполнения операций. Многие отмечают также, что все зависит от того, что имеется в виду под пониманием. Кроме того, в случае отказа считать поведенческий тест подобный тесту Тьюринга достаточным основанием для определения способности к мышлению у нас нет иных оснований считать разумными других людей.

[4] [10] Rey G. What's really going on in searle's 'chinese room' // Philosophical Studies. — 1986. — Vol. 50, no. September. — P. 169–85.

[5] [11] Churchland P. M., Churchland P. S. Could a machine think? // Scienti c American. — 1990. — Vol. 262, no. 1. — P. 32–37.

Заключение

Мысленные эксперименты Тьюринга и Сёрля вызвали оживленные междисциплинарные споры, которые до сих пор не привели к консенсусу по поводу того, являются ли эти аргументы ограничением для наших ожиданий относительно способностей искуственного интеллекта и развития вычислительной теории сознания. Многие исследователи убеждены, что в лучшем случае компьютеры способны имитировать процессы мысления, поскольку из синтаксических операций, которые они осуществляют, не выводится семантика, без которой понимание немыслимо. В то же время кажутся разумными возражения по поводу того, что сознание и мышление как целое невыводимо из относительно простых составляющих, на которых они основаны, и если мозг в процессе мышления ведет себя подобно вычислительной системе, оперирующей символами и нативными представлениями, нет оснований считать, что должным образом запрограммированные машины не будут способны к мышлению.

Вопреки простоте постановки, аргументы поднимают серьезные философские проблемы. Такие вопросы, как природа семантики, ее связь с синтаксисом, внутренними взаимосвязями, биологической природой человека, жизненно важны для понимания природы мышления и сознания и дальнейшего развития искуственного интеллекта, как слабого, так и сильного.

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.

9. Dennett D. C. Toward a cognitive theory of consciousness // Minnesota Studies in the Philosophy of Science. — 1978. — Vol. 9.

10. Rey G. What's really going on in searle's 'chinese room' // Philosophical Studies. — 1986. — Vol. 50, no. September. — P. 169–85.

11. Churchland P. M., Churchland P. S. Could a machine think? // Scientific American. — 1990. — Vol. 262, no. 1. — P. 32–37.