ОБ АВТОРЕ

Кристоф Кох (Christof Koch) — начальник научного отдела и президент Института Аллена по исследованию мозга, научный консультант журнала *Scientific American*.

СОЗНАНИЕ

Пруст среди машин

Уровень компьютерного интеллекта может приблизиться к человеческому уже на нашем веку. Но смогут ли компьютеры осознанно воспринимать окружающий мир?

Кристоф Кох

Стремительно приближается то время, когда мыслительные способности компьютеров достигнут человеческого уровня. Алгоритмы машинного обучения (МО) становятся все более мощными, и мы чувствуем, как они уже дышат нам в затылок. Благодаря стремительному прогрессу в ближайшие десятилетия появятся машины с интеллектом человеческого уровня, способные говорить и размышлять, которые внесут громадный вклад в экономику, политику и в военное дело. Рождение настоящего искусственного интеллекта неизбежно повлияет сильнейшим образом на будущее человечества, в том числе на то, будет ли оно у нас вообще.

Вот две цитаты в качестве наглядного примера:

«С тех пор как в конце 1940-х гг. произошел последний большой прорыв в области искусственного интеллекта, ученые всего мира искали способы его использования для совершенствования технологий, чтобы выйти за рамки того, чего может достичь сегодня самая сложная программа искусственного интеллекта».

«Даже сейчас идут исследования, чтобы выяснить, что новые программы искусственного интеллекта смогут делать, оставаясь в рамках текущего уровня интеллекта. Большинство программ ИИ, созданных до настоящего времени, ограничивались в основном принятием простых решений или выполнением простых операций с относительно небольшими наборами данных».

Эти два абзаца написал *GPT-2*, языковый бот, которого я опробовал прошлым летом. Бот был разработан расположенной в Сан-Франциско организацией *OpenAI*, продвигающей дружественный ИИ. *GPT-2* — это алгоритм МО с дурацкой на первый взгляд задачей: ему дают произвольный начальный текст, а он должен предсказать следующее слово. Сеть не учат «понимать» написанное в сколько-нибудь человеческом смысле. В процессе обучения бот настраивает внутренние связи

Illustration by Gérard Dubois

в своих искусственных нейронных сетях так, чтобы лучше всего предсказывать следующее слово. После обучения на материале 8 млн веб-страниц он содержит более миллиарда соединений, имитирующих синапсы, соединяющие нейроны. Когда я ввел первые несколько предложений статьи, которую вы читаете, алгоритм выдал два абзаца, которые звучали, как попытка первокурсницы, задремавшей на вводной лекции по машинному обучению, вспомнить, о чем шла речь. На выходе имеются все нужные слова, собранные во фразы, — на самом деле, весьма неплохой результат! При повторном вводе того же текста алгоритм выдаст что-нибудь новое.

В будущем такие боты создадут лавину «глубинно поддельных» отзывов о продуктах и новостей, которые усилят загрязнение интернета. Они станут еще одним примером программ, которые выполняют действия, до сих пор считавшиеся исключительно человеческими, — играют в стратегическую игру *StarCraft*, переводят текст, дают личные рекомендации про книги и фильмы, узнают людей на фотографиях и видео.

В дальнейшем потребуется еще много достижений в области машинного обучения, прежде чем алгоритм сможет написать шедевр столь же цельный, как роман Марселя Пруста «В поисках утраченного времени», но это неизбежно случится. Вспомним, что все первые попытки играть в компьютерные игры, переводить и говорить были неуклюжи и не воспринимались всерьез, потому что им явно не хватало мастерства и сноровки. Но с изобретением глубоких нейронных сетей у информационных технологий появилась мощная вычислительная инфраструктура, компьютеры стали непрерывно совершенствоваться, и теперь их достижения уже не кажутся смешными. Как мы уже видели на примере шахмат и покера, современные алгоритмы могут обыграть людей. Когда они это делают, наш первоначальный смех сменяется ужасом. Не вызвали ли мы полезных духов, которых теперь неспособны контролировать, как это сделал ученик чародея в балладе Гете?

Искусственное сознание?

Несмотря на то что среди экспертов нет единого мнения о том, из чего именно состоит интеллект, настоящий или искусственный, большинство согласны, что рано или поздно компьютеры достигнут того, что на жаргоне программистов называется «общий искусственный интеллект» (ОИИ).

Непонятно другое: будет ли ОИИ что-нибудь чувствовать? Смогут ли программируемые компьютеры когда-нибудь обрести сознание?

Под «сознанием», или «субъективным ощущением», я подразумеваю то, что присуще любому переживанию, — например, восхитительный вкус пасты *Nutella*, острое жжение больного зуба, медленное течение времени, когда человек скучает, или чувство бодрости и тревоги непосредственно перед началом соревнований. Говоря в духе философа Томаса Нагеля, система обладает сознанием, если возможно быть этой системой.

Вспомните то неловкое ощущение, когда вы внезапно осознали, что совершили оплошность и ваша шутка воспринята как оскорбление. Смогут ли компьютеры когда-нибудь испытывать подобные эмоции? Когда вы звоните по телефону, минуты идут одна за другой, вы ждете, а искусственный голос нараспев произносит: «Сожалеем, что заставили вас ждать»; действительно ли программное обеспечение чувствует себя виноватым, погружая вас в пекло общения со службой поддержки?

Нет никаких сомнений в том, что наши интеллект и опыт — это неизбежное следствие естественных способностей мозга, а не каких-то сверхъестественных причин. Этот постулат хорошо служил науке, когда люди исследовали мир в течение нескольких последних столетий. Похожий на тофу полуторакилограммовый человеческий мозг — это, безусловно, самый сложный кусок активной материи в известной нам части Вселенной. Но он должен подчиняться тем же физическим законам, которым подчиняются собаки, деревья и звезды. Тут не может быть исключений. Мы пока не до конца понимаем, как работа мозга создает переживания, но мы ощущаем их постоянно: одна группа нейронов активна, когда вы видите цвета, а клетки в другой области коры работают, когда вы в шутливом настроении. Когда нейрохирург стимулирует эти нейроны электродами, люди видят цвета или начинают смеяться. И наоборот, выключение мозга при наркозе устраняет эти ощущения.

Если учитывать широко известные предположения об устройстве сознания, следует ли ожидать, что развитие настоящего искусственного интеллекта приведет к появлению искусственного сознания?

Размышляя над этим вопросом, мы неизбежно приходим к развилке, откуда идут два принципиально разных пути. Дух времени, воплощенный в таких романах и фильмах, как «Бегущий по лезвию», «Она» и «Из машины», решительно ведет нас в сторону предположения, что по-настоящему умные машины будут чувствовать. Они смогут говорить, рассуждать, заниматься самоконтролем и самоанализом. Следовательно, у них будет сознание.

ОСНОВНЫЕ ПОЛОЖЕНИЯ

- В скором времени появятся машины с человеческим уровнем интеллекта.
- Неясно, будет ли у них настоящее сознание.
- Почему? Вряд ли осознанные чувства могут возникнуть даже у самых сложных моделей мозга.

Данное направление наиболее ярко отражено в теории глобального нейронного рабочего пространства (ГНРП), одной из главных научных теорий сознания. Она утверждает, что сознание возникает из-за некоторых особенностей строения мозга.

Ее происхождение можно проследить до систем, использующих принцип «классной доски» в информатике 1970-х гг., когда специализированные программы обращались к общему хранилищу информации, называвшемуся классной доской, или центральным рабочим пространством. Психологи утверждали, что такой же механизм обработки информации существует в мозге и обеспечивает человеческое мышление. У него маленький объем, поэтому только одно ощущение, мысль или воспоминание занимают рабочее пространство в каждый момент времени. Новая информация конкурирует со старой и вытесняет ее.

Сотрудники Коллеж де Франс — специалист по когнитивной нейробиологии Станислас Деан (Stanislas Dehaene) и молекулярный биолог Жан-Пьер Шанже (Jean-Pierre Changeux) — объяснили, как эти идеи соответствуют строению самого внешнего слоя серого вещества головного мозга — коры. Два листа шириной и толщиной с пиццу среднего размера, один справа, а другой слева, сильно смяты и втиснуты в защищающий их череп. Деан и Шанже предположили, что рабочее пространство образовано сетью пирамидных возбуждающих нейронов, связанных с обширными областями коры, в частности префронтальной, височно-теменной и поясной ассоциативными зонами.

Бóльшая часть активности мозга остается привязанной к одному месту и поэтому бессознательной. Например, так работает модуль, контролирующий направление взгляда (мы почти не замечаем, как это делаем), или модуль, регулирующий положение тела. Но когда активность в одной или нескольких областях превышает пороговое значение, например когда кому-то показывают банку с *Nutella*, это вызывает вспышку, волну нервного возбуждения, которое распространяется по всему рабочему пространству, по всему мозгу. Таким образом, эти сигналы становятся доступными для множества дополнительных процессов, таких как язык, планирование, система подкрепления, попадают в буфер кратковременной памяти и получают доступ к долговременной. За счет обширного распространения информации она становится сознательной. Уникальное ощущение от восприятия *Nutella* возникает благодаря пирамидным нейронам, взаимодействующим с областью планирования движений, которая командует схватить ложку и зачерпнуть ореховую пасту. В то же время другие модули передают сигнал об ожидаемом подкреплении в виде большой порции дофамина, поскольку в лакомстве содержится много жира и сахара.

Сознательные ощущения возникают из-за того, что этот алгоритм взаимодействует с сенсорными входами, моторными выходами и внутренними переменными, связанными с памятью, мотивацией и предвкушением. Наличие сознания свидетельствует о том, что происходит обширная обработка. Теория ГНРП полностью соответствует современным идеям о почти бесконечных возможностях вычислительных систем. Сознание — это всего лишь хитрая уловка.

Внутренняя причинная сила

Альтернативный вариант — теория интегрированной информации (ТИИ) — использует более фундаментальный подход для объяснения сознания.

Главный автор проекта — психиатр и нейробиолог из Висконсинского университета в Мадисоне Джулио Тонони (Giulio Tononi). Вместе с другими учеными свой вклад в эту теорию вношу и я. В данной теории все начинается с полученного опыта и активации синапсов, обеспечивающих «ощущение» этого опыта. Интегрированная информация — математическое выражение количества «внутренней причинной силы», которой обладает тот или иной процесс. Возбуждающиеся нейроны, передающие через синапсы сигналы низлежащим клеткам, — это один из механизмов наряду с электронными схемами, состоящими из транзисторов, конденсаторов, резисторов и проводов.

Внутренняя причинная сила — это не какое-то абстрактное понятие, ее можно точно оценить для любой системы. Чем сильнее зависят от текущего состояния системы воспринимаемые входящие сигналы (причины) и исходящие сигналы (следствия), тем большей причинной силой она обладает.

Согласно ТИИ, любой механизм будет иметь сознание, если он обладает внутренней силой и его состояние связано с его прошлым и будущим. Чем больше интегрированная информация системы, обозначаемая греческой буквой Φ («фи»), которая может принимать нулевое или положительное значение, тем больше сознания имеется у системы. Если у чего-то нет внутренней причинной силы, то оно ничего не чувствует, его Φ равно нулю.

Если учитывать неоднородность нейронов коры и сильное перекрывание у них входящих и исходящих связей, объем интегрированной информации в коре огромен. С опорой на эту теорию был создан измеритель сознания, который сейчас тестируют в клинике. С его помощью определяют, находится ли человек в устойчивом вегетативном состоянии или в состоянии минимального сознания, под действием анестезии, или у него синдром «запертого человека», когда сознание присутствует, но человек не может вступать в коммуникацию, или же там внутри уже никого нет. Если оценить причинную силу программируемых цифровых компьютеров на уровне металлических деталей —

транзисторов, проводов и диодов, которые составляют физическую основу для любого вычисления, — получается, что их внутренняя причинная сила и их Φ ничтожно малы. Более того, Φ не зависит от того, какое используется программное обеспечение, происходит ли вычисление суммы налогов или имитация работы мозга.

На самом деле, согласно этой теории, две сети, выполняющие одинаковые операции ввода-вывода, но имеющие разную конфигурацию, могут обладать разным количеством Φ. У одной сети может не быть Φ, а у другой Φ может принимать большое значение. Хотя внешне они кажутся идентичными, одна из них что-то переживает, тогда как ее зомби-двойник не чувствует ничего. Разница скрыта внутри, во внутренней схеме сети. Короче говоря, сознание — это состояние, а не действие.

Разница между этими двумя теориями состоит в том, что ГНРП объясняет сознание через работу человеческого мозга, а ТИИ утверждает, что на самом деле значение имеют внутренние причинные силы мозга.

Различия имеют значение при описании коннектома мозга, то есть определении всех синаптических связей целой нервной системы. Анатомы уже описали коннектом некоторых червей. Они работают над коннектомом дрозофилы и планируют всерьез взяться за коннектом мыши в течение следующего десятилетия. Давайте представим, что в будущем станет возможно сканировать на ультраструктурном уровне весь человеческий мозг со всеми примерно 100 млрд нейронов и квадриллионом синапсов после смерти его хозяина, а затем смоделировать этот орган на каком-нибудь продвинутом компьютере, например квантовом. Если модель достаточно точна, она проснется и станет вести себя как цифровая копия умершего человека — говорить и обращаться к его воспоминаниям, желаниям, страхам и другим чертам личности.

Если теория ГНРП верна и имитации работы мозга достаточно, чтобы получить сознание, смоделированный человек будет обладать сознанием, воплотившись в компьютере. И действительно, загрузка коннектома на сервер, чтобы люди могли жить в цифровой загробной жизни, — обычный ход в научной фантастике.

ТИИ предлагает совершенно другую интерпретацию такой ситуации: модель будет чувствовать себя так же, как программное обеспечение, работающее в японском «умном» унитазе, то есть никак. Она будет действовать как личность, но без каких-либо естественных чувств, словно зомби (правда, без желания питаться человеческой плотью).

Для того чтобы получить сознание, необходимы внутренние причинные силы мозга. Их нельзя смоделировать, они изначально должны быть физически встроены в механизм. Чтобы понять, почему моделирование не работает, задумайтесь, почему никогда не становится мокро при моделировании ливня, или почему астрофизики могут моделировать огромную гравитационную силу черной дыры, не опасаясь, что их поглотит пространство-время, изгибающееся около их компьютера. Ответ: потому что у модели нет причинной силы, чтобы заставить атмосферный пар сконденсироваться в капли или вызвать искривление пространства-времени! Теоретически, однако, можно было бы достичь человеческого уровня сознания, если не ограничиваться моделированием, а создать так называемую нейроморфную технику со структурой, устроенной по образу нервной системы.

Помимо различий в результатах моделирования, есть и другие. ТИИ и ГНРП предсказывают различное положение в коре эпицентра определенных сознательных переживаний, в одном случае он расположен в задней части коры, а в другом — в передней. Это и другие предсказания сейчас проверяются в крупномасштабном совместном проекте шести лабораторий в США, Европе и Китае, получившем только что финансирование в размере $5 млн от Всемирного благотворительного фонда Templton.

Вопрос, могут ли машины иметь сознание, важен и по этическим причинам. Если компьютеры ощущают жизнь с помощью своих собственных чувств, они перестают быть исключительно средством достижения полезных для нас целей. Они становятся самоцельными.

Согласно ГНРП, они превращаются из простых объектов в субъекты, у них появляется их собственное «я». Эта проблема убедительно показана в телевизионных сериалах «Черное зеркало» и «Мир Дикого Запада». Как только когнитивные способности компьютеров начнут соперничать с человеческими, у машин появится непреодолимое стремление отстаивать свои юридические и политические права — право не быть уничтоженными, не подвергаться стиранию памяти, не страдать от боли и разрушения. ТИИ предлагает альтернативный сценарий — что компьютеры останутся всего лишь сверхсложными машинами, призрачными пустыми оболочками, лишенными того, что мы больше всего ценим: ощущения жизни. ■

Перевод: М.С. Багоцкая

ДОПОЛНИТЕЛЬНЫЕ ИСТОЧНИКИ

■ Серл Дж. Разум мозга — компьютерная программа? // ВМН, № 3, 1990.

■ What Is Consciousness, and Could Machines Have It? Stanislas Dehaene, Hakwan Lau and Sid Kouider in Science, Vol. 358, pages 486–492; October 27, 2017.

■ The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed. Christof Koch. MIT Press, 2019.