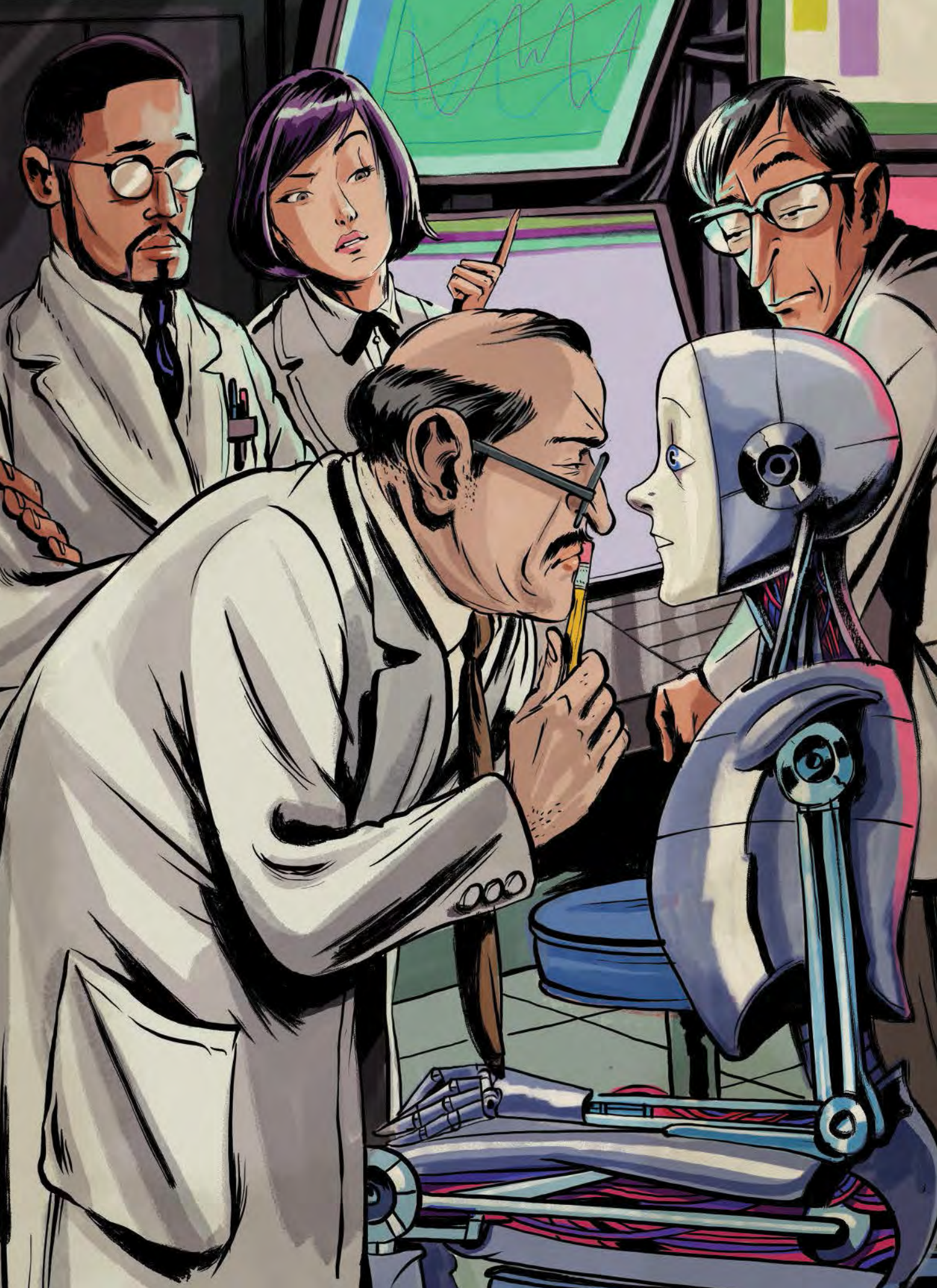


ЧЕЛОВЕК ЛИ Я?

Исследователи нуждаются в новом способе разграничения искусственного и естественного интеллекта

Гэри Маркус

В 1950 г. Алан Тьюринг провел мысленный эксперимент, который с тех пор считается основным критерием для выявления искусственного интеллекта. Он назвал этот эксперимент имитационной игрой, но теперь его именуют тестом Тьюринга. Предвидя появление такого понятия, как виртуальные собеседники (чат-боты — компьютерные программы, которые имитируют поведение человека при общении с другим собеседником), Тьюринг разработал тест, в ходе которого компьютер пытается перехитрить «экзаменатора», притворяясь человеком: рассуждая о прекрасном и умышленно допуская арифметические ошибки. Сегодня обычные люди зачастую рассматривают тест Тьюринга как некий Рубикон, критерий того, действительно ли уже появились умные машины. На самом деле это неправильно: Рубикон можно преодолеть обманным путем, создав машины, способные вводить человека в заблуждение, по крайней мере на непродолжительное время, но такие победы иллюзорны и не идут ни в какое сравнение с интеллектом человека.

ОБ АВТОРЕ

Гэри Маркус (Gary Marcus) — руководитель Лаборатории искусственного интеллекта компании *Uber*, профессор психологии и неврологии Нью-Йоркского университета. Его последняя книга, написанная в соавторстве с Джереми Фриманом (Jeremy Freeman), называется «Будущее мозга» (*The Future of the Brain*, 2014).

Дело в том, что тест Тьюринга пройти слишком просто. Для этого достаточно немного схитрить или изобразить, что ты полный невежда. Как и предполагал сам Тьюринг, нужно притвориться не слишком умным, уходя от большинства вопросов. Если эксперт спросит машину: «Способен ли ты к переживаниям?», ей для достижения своей цели придется солгать. Этот прием вовсе не исключение, а правило. Путь к успеху при прохождении теста Тьюринга — сплошное введение в заблуждение.

Рассмотрим один пример. В 2014 г. чат-бот под ником «Евгений Густман» прошел тест Тьюринга, притворившись 13-летним подростком-весельчаком из Одессы, который никогда не отвечал на вопросы напрямую. Некоторые эксперты поддались на эту уловку (примерно треть от общего числа — достаточно, чтобы считать это небольшой, но все же победой). И что из того? У экспертов было всего несколько минут для общения с экзаменуемым, и, несомненно, при более продолжительном разговоре они вывели бы его на чистую воду. На самом деле «Евгений» не знал — и не понимал — почти ничего. Вот какой диалог вел компьютерщик Скотт Ааронсон (Scott Aaronson) с этим ботом:

Скотт: Что больше, коробка для обуви или гора Эверест?

>> Евгений: Я не могу ответить прямо сейчас. Мне нужно подумать. А кроме того, я забыл спросить, откуда вы...

Скотт: Сколько ног у верблюда?

>> Евгений: Где-то от двух до четырех. Может быть, три? :-))) Между прочим, я до сих пор не знаю, кто вы по специальности. Или я это упустил?

Скотт: Сколько ног у многоножки?

>> Евгений: Всего две, но у чернобыльских мутантов их может быть до пяти. Я знаю, вы хотите подловить меня.

Скотт: Нет, мне нужно выяснить, не чат-бот ли вы. Пожалуйста, ответьте прямо на вопрос: сколько ног у муравья?

>> Евгений: Где-то от двух до четырех. Может быть, три? :-))) О, какой плодотворный разговор ;-)

Исходный тест Тьюринга не выдержал испытания временем.

Вскоре после того как история с «Евгением» попала на страницы печатных изданий, я предложил альтернативный тест, разработанный для того, чтобы повысить уровень естественного интеллекта, вместо того чтобы довольствоваться сомнительными отговорками собеседника. В очередном посте в блоге *New Yorker* я высказал мысль, что на смену старому тесту Тьюринга должен прийти новый, современный — «Тест Тьюринга XXI в.».

Моей целью, как я тогда писал, было «создание компьютерной программы, которая могла бы, посмотрев любую произвольную телепередачу или видеоролик в *YouTube*, ответить на вопросы об их содержании». Я хотел исключить возможные уловки и выяснить, способны ли машины на самом деле осознавать то, что они видят или слышат. Создание программ, наделяющих компьютеры «чувством юмора», вряд ли приблизит нас к настоящему искусственному интеллекту, но это поможет проникнуть более глубоко в то, как они воспринимают мир.

Ознакомившись с моими соображениями, Франческа Росси (Francesca Rossi), в то время председатель Международной объединенной конференции по искусственному интеллекту *(IJCAI)*, предложила нам сотрудничество с целью модернизации

ОСНОВНЫЕ ПОЛОЖЕНИЯ

- Долгое время считалось, что «имитационная игра» Алана Тьюринга, в которой машина пыталась убедить экзаменатора, что она человек, — это оптимальный тест для выявления искусственного интеллекта (ИИ).
- Но тест Тьюринга не выдержал испытания временем. Чтобы его пройти, нужна скорее хитрость, чем истинный интеллект. Специалисты в области ИИ утверждают, что пришла пора заменить тест Тьюринга набором тестов, которые будут оценивать интеллектуальные способности машины с самых разных точек зрения.
- По-настоящему умная машина должна понимать смысл неоднозначных высказываний, уметь собирать разобранную и упакованную мебель, быть способной пройти тест по научным предметам за четвертый класс и многое другое. Сложность этих тестов подчеркивает тот факт, что, если говорить по существу, достижение машиной уровня интеллекта человека остается делом далекого будущего.

Новые тесты Тьюринга

Исследователи в области ИИ разрабатывают разнообразные тесты, которые должны прийти на смену 67-летней «имитационной игре» Алана Тьюринга. Здесь мы рассмотрим четыре из них.

Джон Павлус

ТЕСТ 01

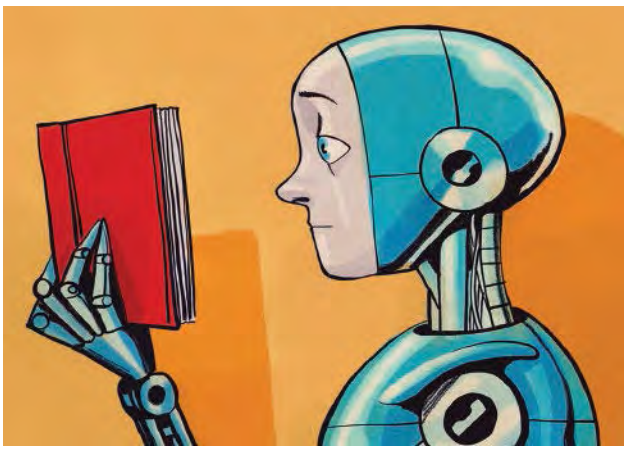


Схема Винограда

Названная в честь первопроходца в области ИИ исследователя Терри Винограда, схема Винограда *(Winograd Schema Challenge)* — это простой, но неоднозначно сформулированный на обычном языке вопрос. Для правильного ответа требуется «осознанное» понимание того, как именно в реальном мире взаимосвязаны люди, предметы и культурные нормы. В первой схеме Винограда, созданной в 1971 г., сначала предлагается некая ситуация («Члены городского совета отказали активистам в проведении митинга, потому что они боялись насилия»), а затем задается простой вопрос («Кто боялся насилия?»). Эта задача известна под названием устранения неоднозначности местоимений *(pronoun disambiguation problem, PDP)*: в данном случае неоднозначность относится к слову «они». Однако схемы Винограда немного более хитроумны, чем многие *PDP*, поскольку смысл предложения можно обратить, заменив лишь одно слово. (Например: «Члены городского совета отказали активистам в проведении митинга, потому что они пропагандировали насилие».) Чтобы разобраться, в чем тут дело, обычные люди используют здравый смысл или общеизвестные представления об обычных взаимоотношениях между городскими властями и демонстрантами. В этом тесте применяется простейший раунд *PDP*, с тем чтобы отсеять системы с наименьшим уровнем интеллекта; успешно прошедшие его допускаются к тестированию в рамках полноразмерных схем Винограда.

Плюсы. Поскольку схемы Винограда опираются на данные, надежного доступа к которым тестируемые компьютеры не имеют, этот тест «гуглоустойчив» (то есть экзаменуемый не способен найти прямой ответ в *Google*).

Минусы. Число используемых схем сравнительно невелико. «Их не так просто придумать», — говорит Эрнест Дэвис, профессор вычислительной техники и информатики из Нью-Йоркского университета.

Уровень сложности. Высокий. В 2016 г. четыре системы соревновались друг с другом, отвечая на вопросы 60 схем Винограда. Победитель смог правильно ответить лишь на 58% из них — намного ниже порога в 90%, который эксперты определили как проходной балл.

Для чего это нужно. Для того чтобы отличать осмысление от его имитации. «*Siri*, электронный персональный помощник от компании *Apple*, не осознает значения местоимений и не способен устранять неоднозначность толкования», — поясняет Леора Моргенштерн, исследователь из компании *Leidos*, которая вместе с Дэвисом занималась разработкой схемы Винограда. Это означает, что «вы фактически не можете вести диалог с системой, поскольку постоянно ссылаетесь на что-то, о чем шла речь в предыдущем разговоре».

ТЕСТ 02

Типовые испытания умных электронных устройств

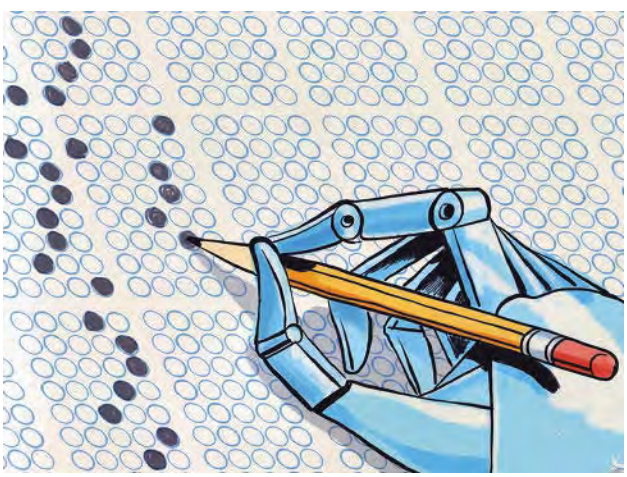
Устройствам, обладающим искусственным интеллектом, можно было бы без каких-либо поблажек предложить типовые письменные тесты, которые проходят ученики начальных и средних классов школы. Этот метод позволяет оценивать способность машины связывать факты воедино с помощью новых способов, основанных на осмысленном понимании. Так же как и «имитационная игра» Тьюринга, эта схема гениально проста. Возьмите любое достаточно строгое типовое исследование (например, многократную выборку государственного экзамена за четвертый класс по точным наукам в различных частях штата Нью-Йорк), научите машину усваивать тестовый материал (например, с помощью обработки естественного языка и компьютерного зрения) — и вперед.

Плюсы. Универсальность и практичность. В отличие от схем Винограда, стандартизованный тестовый материал недорог и имеется в избытке. И поскольку ни один из этих материалов не адаптирован и не обработан предварительно, что могло бы дать машинам какое-либо преимущество, для ответа на вопросы требуются энциклопедические знания и здравый смысл, поэтому правильных ответов дается гораздо меньше.

Минусы. Не настолько «гуглоустойчив», как схемы Винограда и некоторые люди; успешное прохождение типового теста не обязательно подразумевает наличие «естественного» интеллекта.

Уровень сложности. Умеренно высокий. Система под названием *Aristo*, созданная в Институте искусственного интеллекта Пола Аллена, на экзамене по точным наукам за четвертый класс достигла результата в среднем на уровне 75%, чего ни разу не удавалось показать до сих пор. Но это был только набор многочисленных вопросов без всяких диаграмм. «На сегодня ни одной системе не удается даже близко подойти к успешному прохождению экзаменов за четвертый класс по точным наукам», — написали исследователи из института Аллена в статье, опубликованной в журнале *AI Magazine*.

Для чего это нужно. Для проверки достоверности результатов. «Мы видим, что ни одна программа не может подняться выше 60-процентного уровня на экзамене по точным наукам за восьмой класс, но в то же время мы читаем в новостях, что суперкомпьютер *Watson*, созданный компанией *IBM*, окончил медицинский институт и будет использоваться в качестве врача-диагноста, — говорит Орен Этциони (Oren Etzioni), главный исполнительный директор Института искусственного интеллекта Пола Аллена. — Либо *IBM* совершили грандиозный прорыв, либо, возможно, они немного опережают события».

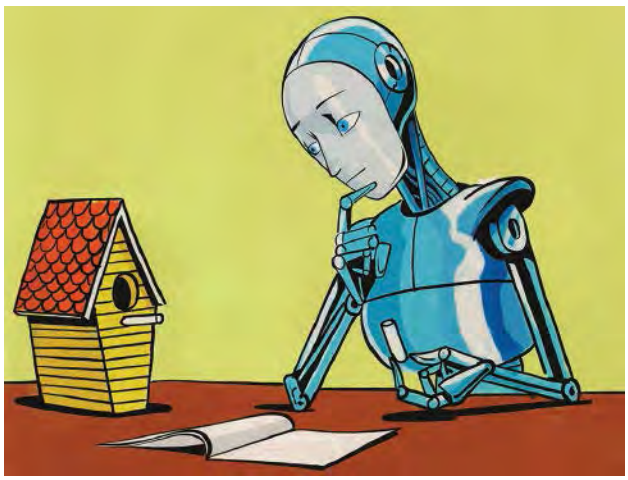


ТЕСТ 03

Новые тесты Тьюринга

Продолжение

ТЕСТ 04



Материально воплощенный тест Тьюринга

Многие тесты для оценки уровня интеллекта машин направлены на выявление их способностей к распознаванию. Этот тест больше всего напоминает урок труда: искусственный интеллект должен осмысленно манипулировать реальными объектами. Испытание предусматривает два этапа. На этапе строительства искусственный интеллект в материальном воплощении — по существу, это робот — пытается собрать конструкцию из груды деталей, используя вербальные, написанные и нарисованные инструкции (что-то вроде сборки мебели из *IKEA*). На этапе исследования от робота потребуется найти решения нескольких допускающих разные толкования, но с возрастающей степенью креативности задач, используя детали детского конструктора (например, «построить стену», «построить дом», «пристроить к дому гараж»). Кульминацией каждого этапа будет решение коммуникативной задачи, где роботу понадобится «обосновать» свои действия. Этот тест могут проходить как отдельные роботы, так и группы роботов или роботы в сотрудничестве с людьми.

Плюсы. Этот тест объединяет аспекты интеллектуальной деятельности в реальном мире — в частности, восприятие и действие, что ранее не учитывалось или недооценивалось. Кроме того, тест практически невозможно перехитрить: «Я не знаю, как это можно сделать без подсказки», — говорит Чарлз Ортиз из компании *Nuance Communications*.

Минусы. Трудоемкий, длительный, плохо поддается автоматизации, если машины работают не в виртуальной реальности. И даже в этом случае «робототехники скажут, что это [виртуальная реальность] — всего лишь аппроксимация, — говорит Ортиз. — В реальном мире, когда вы берете предмет, он может выскользнуть из рук или может, например, подуть ветер, который как-то повлияет на ситуацию. В виртуальном мире сложно учесть все эти нюансы».

Уровень сложности. Научная фантастика. Материализованный искусственный интеллект, который мог бы правильно манипулировать предметами и связно объяснять свои действия, то есть, по существу, вести себя как дроид из эпопеи «Звездные войны», сейчас далеко за пределами наших возможностей. «Выполнение подобных заданий так, как это делают обычные дети, пока невероятно сложно», — говорит Ортиз.

Для чего это нужно. Для понимания пути к интеграции четырех составляющих ИИ — восприятия, действия, распознавания и языка, которым, как правило, посвящены отдельные исследовательские программы.

I-Athlon

Множество существующих сегодня частично или полностью автоматизированных тестов на ИИ предлагают кратко изложить содержание аудиофайла и сюжет видеофрагмента, с ходу выполнить перевод и решить другие задачи. Цель теста — получить объективную оценку уровня интеллекта. Автоматизация тестирования и оценки (то есть осуществление их без участия человека) — отличительная особенность этой схемы. Отстранение человека от процесса оценки уровня интеллекта машины может показаться странным, но, как говорит Мюррей Кэмпбелл (Murray Campbell), исследователь в области ИИ из компании *IBM* (и член команды, разработавшей шахматный суперкомпьютер *Deep Blue*), это необходимо для обеспечения оперативности и воспроизводимости тестирования. Как замечает Кэмпбелл, определение уровня умственных способностей устройства с искусственным интеллектом на основе алгоритмов избавляет от необходимости полагаться на разум человека — «со всеми его когнитивными погрешностями и субъективностью» — как на неоспоримый критерий.

Плюсы. Объективность, по крайней мере теоретически. Как вести счет в каждом тесте и оценивать результаты, прежде решали *I-Athlon*-эксперты, теперь это будет делать компьютер, причем беспристрастно. Судить о результатах следует так же четко и объективно, как это делает фотофиниш на Олимпийских играх. Разнообразие тестов также поможет идентифицировать — в соответствии с определением экспертов из *IBM* — «истинно умные системы».

Минусы. Потенциальная туманность. *I-Athlon*-алгоритмы могут высоко оценить системы, обладающие ИИ, которые работают не совсем понятным для исследователей образом. «Вполне возможно, что некоторые решения продвинутых умных машин будет очень сложно объяснить [человеку] коротко и ясно», — признает Кэмпбелл. Эта так называемая проблема черного ящика уже стоит перед исследователями, работающими со сверточными нейронными сетями.

Уровень сложности. Зависит от обстоятельств. Современные системы могут показать хорошие результаты в некоторых *I-Athlon*-тестах, например тестах на распознавание изображений или языковых переводах. Другие аспекты, такие как изложение содержания видеоролика или построение диаграммы по словесному описанию, пока относятся к области научной фантастики.

Для чего это нужно. Для уменьшения влияния когнитивной погрешности человека на оценку умственных способностей машины и квантификацию — а не просто определение — ее качества.

теста Тьюринга. Вместе мы заручились поддержкой Мануэлы Велосо (Manuela Veloso), специалиста в области робототехники из Университета Карнеги — Меллона и бывшего президента Ассоциации по развитию искусственного интеллекта, и втроем предприняли мозговой штурм. Вначале мы сосредоточились на разработке отдельного теста, который мог бы заменить детище Тьюринга, но вскоре обратились к идее создания множества тестов, решив, что поскольку нет универсального теста для определения уровня физической подготовки спортсменов, не может быть такового и для оценки уровня интеллекта.

Кроме того, мы решили привлечь как можно больше экспертов в области искусственного интеллекта (ИИ) и собрали в январе 2015 г. в Остине, штат Техас, примерно 50 ведущих специалистов, чтобы обсудить проблему модернизации теста Тьюринга. Посвятив целый день докладам и обсуждениям, мы сошлись на идее конкуренции нескольких тестов.

Один из них, тест *Winograd Schema Challenge*, названный в честь основоположника принципов построения ИИ Терри Винограда (Terry Winograd), наставника разработчиков и основателей поисковой системы *Google* Ларри Пейджа (Larry Page) и Сергея Брина, предлагает подвергать машины испытанию, где важны как понимание языка, так и здравый смысл. Всякий, кто хоть раз пытался обучить электронную машину понимать человеческий язык, очень быстро осознавал, что смысл практически каждого предложения неоднозначен и зачастую его можно интерпретировать несколькими способами. Наш мозг так хорошо приспособлен к пониманию языка, что обычно мы этого не замечаем. Возьмем, например, фразу «Большой мяч пробил стол, потому что он был сделан из пенопласта». Строго говоря, предложение неоднозначно: местоимение «он» может относиться как к слову «стол», так и к слову «мяч». Но любой человек понимает, что «он» — это стол. Однако здесь требуются как понимание языка, так и познания в области материаловедения — а это уже слишком сложно для машины. Трое экспертов, Эктор Левеск (Hector Levesque), Эрнест Дэвис (Ernest Davis) и Леора Моргенштерн (Leora Morgenstern), уже разработали тест, касающийся таких предложений, и *Nuance Communications*, компания, занимающаяся методами распознавания речи, выделила $25 тыс. наличными в качестве приза алгоритму, который пройдет этот тест.

Мы надеемся использовать и другие тесты, например *Comprehension Challenge*, проверяющий способность «умных» машин воспринимать изображения, видео- и аудиозаписи, а также текстовую информацию, был бы здесь вполне уместен. Чарлз Ортиз — мл. (Charles Ortiz, Jr.), руководитель Лаборатории искусственного интеллекта и обработки естественного языка в корпорации *Nuance Communications*, предложил использовать *Comprehension Challenge* для проверки способности к восприятию и физическому действию — две важные составляющие разумного поведения, которые никак не отражены в тесте Тьюринга. А Питер Кларк (Peter Clark) из Института искусственного интеллекта Пола Аллена выдвинул идею использовать для машин те же стандартные тесты по научным и другим предметам, которые проходят школьники.

Помимо самих тестов участники конференции обсудили основные критерии, которым должен соответствовать безупречный тест. Так, Гурудут Банавар (Guruduth Banavar) с коллегами из *IBM* отметил, что подобные тесты должны создаваться самими компьютерами. Стюарт Шибер (Stuart Shieber) из Гарвардского университета подчеркнул важность прозрачности: если тесты будут широко использоваться, награды должны получать только открытые — доступные всему ИИ-сообществу — и воспроизводимые системы.

Когда машины смогут достичь уровня установленных нами требований, не знает никто. Однако люди уже всерьез воспринимают некоторые из этих тестов, что может иметь практическое значение. Так, робот, который прошел тест *Construction Challenge*, мог бы, например, возводить временные кампусы для переселенцев на Земле или на далеких планетах. Машина, способная пройти тест *Winograd Schema Challenge* и сдать экзамен по биологии за четвертый класс, приблизила бы нас к реализации мечты о создании компьютеров, которые интегрируют большие массивы медицинских данных, что, вероятно, будет одним из важных шагов в лечении онкологических больных или дешифровке работы мозга. Искусственный интеллект, как и любая область науки, должен иметь четкие цели. Тест Тьюринга был прекрасным началом, но теперь пришло время для решения задач нового поколения. ■

Перевод: С.Э. Шафрановский

ДОПОЛНИТЕЛЬНЫЕ ИСТОЧНИКИ

- Черчленд П., Черчленд П. Может ли машина мыслить? // ВМН, № 3, 1990.
- Computing Machinery and Intelligence. A.M. Turing in Mind, Vol. 59, No. 235, pages 433-460; October 1950.
- What Comes after the Turing Test? Gary Marcus in New Yorker. Опубликовано онлайн 09.06.2014: www.newyorker.com/tech/elements/what-comes-after-the-turing-test
- Beyond the Turing Test. Special issue of AI Magazine, Vol. 37, No. 1; Spring 2016.
- Тест Winograd Schema Challenge: http://commonsensereasoning.org/winograd.html