Архитекторы интеллекта

Вся правда об искусственном интеллекте от его создателей

Иошуа Бенджио
Стюарт Рассел
Джеффри Хинтон
Ник Бостром
Ян Лекун
Фей-Фей Ли
Демис Хассабис
Эндрю Ын
Рана эль Калиуби
Рэймонд Курцвейл
Даниэла Рус
Джеймс Маника
Гари Маркус
Барбара Грош
Джуда Перл
Джефф Дин
Дафна Коллер
Дэвид Ферруччи
Родни Брукс
Синтия Бризил
Джошуа Тененбаум
Орен Этциони
Брайан Джонсон

Автор бестселлеров New York Times

Мартин Форд

Мартин Форд

Архитекторы интеллекта. Вся правда об искусственном интеллекте от его создателей

«Питер»
2018

УДК 004.8
ББК 32.813

Форд М.

Архитекторы интеллекта. Вся правда об искусственном интеллекте от его создателей / М. Форд — «Питер», 2018

ISBN 978-5-4461-1254-8

Искусственный интеллект (ИИ) быстро переходит из области научной фантастики в повседневную жизнь. Современные устройства распознают человеческую речь, способны отвечать на вопросы и выполнять машинный перевод. В самых разных областях, от управления беспилотным автомобилем до диагностирования рака, применяются алгоритмы распознавания объектов на базе ИИ, возможности которых превосходят человеческие. Крупные медиакомпании используют роботизированную журналистику, создающую из собранных данных статьи, подобные авторским. Очевидно, что ИИ готов стать по-настоящему универсальной технологией, такой как электричество. Какие подходы и технологии считаются наиболее перспективными? Какие крупные открытия возможны в ближайшие годы? Можно ли создать по-настоящему мыслящую машину или ИИ, сравнимый с человеческим, и как скоро? Какие риски и угрозы связаны с ИИ и как их избежать? Вызовет ли ИИ хаос в экономике и на рынке труда? Смогут ли суперинтеллектуальные машины выйти из-под контроля человека и превратиться в реальную угрозу? Разумеется, предсказать будущее невозможно. Тем не менее эксперты знают о текущем состоянии технологий, а также об инновациях ближайшего будущего больше, чем кто бы то ни было. Вас ждут блестящие встречи с такими признанными авторитетами, как Р. Курцвейл, Д. Хассабис, Дж. Хинтон, Р. Брукс и многими другими. В формате PDF A4 сохранен издательский макет книги.

УДК 004.8
ББК 32.813

ISBN 978-5-4461-1254-8

Содержание

Мартин Форд
Архитекторы интеллекта. Вся правда об искусственном интеллекте от его создателей

Посвящается Сяо-Сяо, Элейн, Колину и Тристану

Вступление

Мартин Форд

Футуролог, эксперт в области искусственного интеллекта, консультант по расчетам индекса робототехники Rise Of The Robots B Societe Generale, предприниматель из кремниевой долины, член совета директоров и инвесторов компании Genesis Systems

Мартин Форд – автор книг «Роботы наступают: развитие технологий и будущее без работы»[1] *(отмечена премией Financial Times & McKinsey Business Book of the Year, переведена более чем на 20 языков) и «Технологии, которые изменят мир»*[2]*. Писал о технологиях будущего для The New York Times, Fortune, Forbes, The Atlantic, The Washington Post, Harvard Business Review, The Guardian и The Financial Times. Выступал на многочисленных радио- и телешоу, в том числе на NPR, CNBC, CNN, MSNBC и PBS. Часто делает доклады о влиянии ИИ на экономику, рынок труда и общество будущего (наиболее известно выступление на конференции TED 2017 г.). Получил степень бакалавра в области вычислительной техники в Мичиганском университете в Анн-Арборе и степень магистра бизнеса в Калифорнийском университете в Лос-Анджелесе (UCLA). В компании Genesis Systems участвует в разработке автоматизированных систем с автономным питанием, генерирующих воду непосредственно из воздуха.*

Искусственный интеллект (ИИ) быстро переходит из области научной фантастики в повседневную жизнь. Современные устройства понимают человеческую речь, способны отвечать на вопросы и выполнять машинный перевод. В самых разных областях, от управления беспилотным автомобилем до диагностирования рака, применяются алгоритмы распознавания объектов на базе ИИ, возможности которых превосходят человеческие. Крупные медиакомпании используют роботизированную журналистику, создающую из собранных данных статьи, подобные авторским. Очевидно, что ИИ готов стать одним из важнейших факторов, формирующих наш мир, являясь по-настоящему универсальной технологией, такой как электричество.

В последние годы в СМИ широко освещаются достижения в области ИИ. Бесчисленные статьи, книги, документальные фильмы и телепрограммы предсказывают новую эру, мешая в одну кучу анализ фактических данных с ажиотажем, спекуляциями и даже нагнетанием паники. Говорят, что через несколько лет дороги полностью захватят беспилотные автомобили, оставив без работы водителей грузовиков и такси. В некоторых алгоритмах машинного обучения обнаружили признаки дискриминации по расовому и половому признакам. Неясно, как повлияет на конфиденциальность распознавание лиц. Роботы могут стать оружием. А обладающие интеллектом машины представляют угрозу существованию человечества. Свое мнение озвучивают многие общественные деятели, которые не являются экспертами в сфере ИИ. Радикальнее всего выступил Илон Маск, заявивший, что разработки ИИ сродни призыву демонов и опаснее ядерного оружия. Даже Генри Киссинджер и Стивен Хокинг публиковали мрачные прогнозы.

Поэтому мне хотелось бы рассказать о том, что такое ИИ, и осветить связанные с ним возможности и риски. Для этого я провел серию интервью с выдающимися учеными и предпринимателями, занимающимися ИИ. Многие из них лично повлияли на трансформацию окружающего нас мира; другие – основали компании, которые расширяют границы ИИ, робототехники и машинного обучения.

---

[1] *Форд М.* Роботы наступают: развитие технологий и будущее без работы / Пер. с англ. С. Чернина. – М.: Альпина нонфикшн, 2016. – 429 с.: ил. – (Серия «Искусственный интеллект»).

[2] *Форд М.* Технологии, которые изменят мир / Пер. с англ. А. Кардаш. – М.: Манн, Иванов и Фербер, 2014. – 268 с.

Разумеется, сформированный мной список субъективен – в развитии ИИ участвует множество профессионалов. Но я уверен, что почти любой человек, обладающий глубокими знаниями в этой области, поддержит мой выбор. Всех этих людей можно без преувеличения назвать творцами ИИ, приближающими начало новой научно-технической революции.

В интервью я старался задать наиболее острые вопросы, появившиеся в процессе развития ИИ. Какие подходы и технологии считаются наиболее перспективными? Какие крупные открытия возможны в ближайшие годы? Можно ли создать по-настоящему мыслящую машину или ИИ, сравнимый с человеческим, и как скоро? Какие риски и угрозы связаны с ИИ и как их избежать? Требуется ли для этой области государственное регулирование? Вызовет ли ИИ хаос в экономике и на рынке труда? Смогут ли суперинтеллектуальные машины выйти из-под контроля человека и превратиться в реальную угрозу? Нужно ли беспокоиться о «гонке вооружений» в области ИИ?

Разумеется, предсказать будущее невозможно. Тем не менее эксперты знают о текущем состоянии технологий, а также об инновациях ближайшего будущего больше, чем кто бы то ни было. Поэтому их мысли и мнения заслуживают внимания. Помимо ИИ, мы обсудили образование, карьеру и исследовательские интересы, поэтому чтение будет увлекательным и вдохновляющим.

Искусственный интеллект – это широкая область исследований, сопряженная с множеством дополнительных дисциплин. Многие из моих собеседников совмещали работу в нескольких областях. Сейчас я кратко расскажу, как опрошенные относятся к наиболее важным инновациям в исследованиях ИИ и задачам будущего. Основная информация о каждом из них будет приведена в начале соответствующего интервью.

Подавляющее большинство достижений сферы ИИ последнего десятилетия – от распознавания лиц до машинного перевода и победы в игре го – основаны на технологии глубокого обучения, или глубоких нейронных сетей. Искусственные нейронные сети, в которых программно эмулируется структура и взаимодействие нейронов головного мозга, появились примерно в 1950-х гг. Простые версии этих сетей могли решать элементарные задачи по распознаванию объектов на изображениях, что сначала вызывало сильный энтузиазм. Однако к 1960 г., частично из-за критики Марвина Минского – одного из пионеров ИИ, – нейронные сети потеряли популярность, а им на смену пришли другие подходы.

В течение примерно 20 лет, начиная с 1980-х гг., небольшая группа исследователей продолжала верить в технологию нейронных сетей и продвигать ее. Среди них выделялись Джеффри Хинтон (Geoffrey Hinton), Иошуа Бенджио (Yoshua Bengio) и Ян Лекун (Yann LeCun). Они не только внесли вклад в лежащую в основе глубокого обучения математическую теорию, но и первыми стали продвигать технологию «глубоких» сетей с несколькими слоями искусственных нейронов. Им удалось донести идею нейронных сетей до времен экспоненциального роста вычислительных мощностей и увеличения объема доступных данных. В 2012 г. команда аспирантов Хинтона из Университета Торонто победила в конкурсе по распознаванию объектов на изображениях.

После этого события глубокое обучение стало общедоступным. Большинство крупных технологических компаний – Google, Facebook, Microsoft, Amazon, Apple, Baidu и Tencent – инвестировали огромные суммы в новую технологию, чтобы использовать ее в своем бизнесе. Разработчики микропроцессорных и графических чипов (GPU), такие как NVIDIA и Intel, переорганизовали бизнес под создание оборудования, оптимизированного для нейронных сетей. Именно глубокое обучение сегодня раскрывает сферу ИИ.

Такие ученые, как Эндрю Ын (Andrew Ng), Фей-Фей Ли (Fei-Fei Li), Джефф Дин (Jeff Dean) и Демис Хассабис (Demis Hassabis), используют современные нейронные сети в таких областях, как поисковые системы, компьютерное зрение, беспилотные автомобили и универ-

сальный ИИ. Это признанные лидеры в области преподавания, управления и предпринимательства на базе технологии нейронных сетей.

Однако глубокое обучение подвергается критике. Ряд ученых считают его «одним из инструментов в наборе», утверждая, что для дальнейшего прогресса нужны идеи из других областей. Барбара Грош (Barbara Grosz) и Дэвид Ферруччи (David Ferrucci) занимаются проблемами понимания естественного языка. Гари Маркус (Gary Marcus) и Джош Тененбаум (Josh Tenenbaum) изучают человеческое познание. Орен Этциони (Oren Etzioni), Стюарт Рассел (Stuart Russell) и Дафна Коллер (Daphne Koller) специализируются на вероятностных методах. Джуда Перл (Judea Pearl) за работу по вероятностным (или байесовским) подходам к ИИ и машинному обучению получил премию Тьюринга.

Сфера робототехники также развивается благодаря таким ученым, как Родни Брукс (Rodney Brooks), Даниэла Рус (Daniela Rus) и Синтия Бризил (Cynthia Breazeal). Бризил вместе с Раной эль Калиуби (Rana El-Kaliouby) – первопроходцы в построении систем, умеющих распознавать эмоции, реагировать на них и вступать в социальные взаимодействия с людьми. Брайан Джонсон (Bryan Johnson) основал компанию *Kernel*, направляющую технологии ИИ в развитие человека.

По моему мнению, особый интерес представляют три основные темы, поэтому они будут рассматриваться в каждом интервью. Первая касается автоматизации человеческого труда, ведущей к росту безработицы. Глубже всего эту тему раскрыл Джеймс Маника (James Manyika) – глава Глобального института McKinsey (MGI), где активно исследуется влияние технологий на рынок труда.

Второй вопрос, который я задавал всем, касается ИИ, сравнимого с человеческим. Это так называемый сильный ИИ (artificial general intelligence, AGI), который был недостижимой мечтой. Демис Хассабис рассказал, что предпринимается в компании DeepMind, которая является крупнейшей и наиболее финансируемой инициативой по исследованиям сильного ИИ. Дэвид Ферруччи, руководивший разработкой суперкомпьютера IBM Watson, – генеральный директор стартапа Elemental Cognition, – описал создание сильного ИИ путем эффективного применения понимания языка. Важные идеи высказал и Рэймонд Курцвейл (Raymond Kurzweil) – автор книги *Singularity is Near* («Сингулярность уже близка»), в настоящее время руководящий проектом Google, связанным с обработкой естественного языка.

Всем я задал вопрос: «К какому году с вероятностью 50 % будет создан ИИ уровня человеческого?» Большинство предпочло поделиться своими предположениями анонимно. Двое из опрошенных выразили желание официально поделиться своей точкой зрения. Результаты этого опроса приведены в конце книги. Вы увидите, как мнения по важным темам зачастую кардинально расходятся, что представляет собой один из наиболее интересных аспектов данной книги.

Третья обсуждаемая тема связана с последствиями прогресса в области ИИ, ожидаемыми как в ближайшем, так и в отдаленном будущем. Становится очевидной проблема уязвимости взаимосвязанных автономных систем к атакам через интернет. Также выявлена предрасположенность алгоритмов машинного обучения к предвзятости по расовому и половому признакам. Многие из моих собеседников подчеркнули важность решения этой проблемы и рассказали об исследованиях в этой области. Некоторые дали оптимистический прогноз, предположив, что ИИ поможет нам избавиться от предвзятости и дискриминации.

Многих волнует опасность появления полностью автономного оружия. В сообществе исследователей ИИ существует мнение, что роботы или дроны, способные убивать без контроля человека, в конечном итоге могут стать не менее опасными, чем биологическое или химическое оружие. В июле 2018 г. более 160 компаний и 2400 исследователей (среди которых

есть мои собеседники) подписали соглашение о запрете производства смертоносных алгоритмов[3].

Более отдаленной и умозрительной является проблема несоответствия собственных целей сильного ИИ с желаниями человека. Этой темы касались почти все мои собеседники. Чтобы адекватно и рационально осветить эту проблему, я поговорил с Ником Бостромом (Nick Bostrom) из оксфордского Института будущего человечества (Future of humanity institute, FHI) – автором бестселлера «Искусственный интеллект: этапы, угрозы, стратегии»[4], в котором тщательно рассматриваются потенциальные риски, связанные с машинами, интеллектуально превосходящими человека.

[3] https://futureoflife.org/lethal-autonomous-weapons-pledge/

[4] *Бостром Н.* Искусственный интеллект: этапы, угрозы, стратегии / Пер. с англ. С. Филина. – М.: Манн, Иванов и Фербер, 2016. – 490 с.: ил.

Создатели интеллекта

Интервью для этой книги проводились с февраля по август 2018 г. Практически все они длились не меньше часа, а некоторые существенно дольше. Записанные, транскрибированные, а затем отредактированные командой издательства Packt тексты я дал своим собеседникам на проверку. Уверен, что книга верно отражает мысли респондентов.

Эксперты, с которыми я общался, имеют разное происхождение и сотрудничают с разными компаниями. Но вы быстро обнаружите, насколько сильно влияние Google на сообщества, связанные с ИИ. Из двадцати трех специалистов у семи есть или были связи с Google или холдингом Alphabet. Много талантливых людей работает в Массачусетском технологическом институте (MIT) и Стэнфорде. Джеффри Хинтон и Иошуа Бенджио представляют университеты Торонто и Монреаля соответственно, а правительство Канады ведет четкую промышленную политику, ориентированную на робототехнику и ИИ. В Соединенных Штатах работали 19 из 23 опрошенных, но больше половины из них родились за пределами США: в Австралии, Китае, Египте, Франции, Израиле, Родезии (ныне Зимбабве), Румынии и Великобритании. Это ярко иллюстрирует роль иммиграции квалифицированных кадров в технологическом лидерстве США.

Проводя интервью, я все время помнил, что книгу будут читать самые разные люди, от специалистов по теории вычислительных машин и систем до менеджеров и инвесторов. Но самая важная часть аудитории – молодые люди, которые могут задуматься о карьере в области ИИ. Сейчас в ней наблюдается дефицит кадров, особенно специалистов с навыками глубокого обучения, что дает возможность для хорошего карьерного роста. В настоящее время прилагаются усилия по привлечению в отрасль талантливых специалистов, и уже широко признается необходимость профессиональной интеграции.

Около четверти опрошенных мной – женщины. Здесь их доля выше, чем в сфере ИИ и машинного обучения в целом. Согласно недавним исследованиям, женщины составляют примерно 12 % ведущих сотрудников в области машинного обучения[5]. В процессе интервью многие подчеркивали необходимость увеличения доли как женщин, так и представителей меньшинств.

Одна из моих собеседниц уделяет особое внимание многообразию в области ИИ. Фей-Фей Ли из Стэнфорда – соучредитель AI4ALL[6], устраивающей летние учебные лагеря для старшеклассников из мало представленных в этой сфере групп. AI4ALL получила поддержку отрасли, а также грант от Google, и теперь такие программы проводятся в шести американских университетах. В этом направлении еще многое предстоит сделать, но основания для оптимистических прогнозов уже есть.

Хотя книга рассчитана на широкий круг читателей, в тексте будут встречаться специальные понятия и термины. Если вы ранее ничего не знали об ИИ, то я рад, что вас познакомят с ним ведущие специалисты, и рекомендую вам начать с краткого словаря, приведенного ниже. В интервью Стюарта Рассела – соавтора ведущего учебника по ИИ – вы найдете объяснение ключевых концепций области.

Возможность взять эти интервью была для меня честью. Надеюсь, вы тоже увидите в моих собеседниках вдумчивость, умение рассказывать и глубокую приверженность идее работы на благо человечества. Чего в книге нет, так это единодушия. Она наполнена разнообразными, зачастую резко противоречивыми представлениями, мнениями и прогнозами. Понятно только одно: ИИ – широко открытое пространство. Можно строить предположения о природе буду-

[5] https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance

[6] http://ai-4-all.org/

щих инноваций, скорости их появления и конкретных вариантах их применения. Именно из-за этой комбинации потенциальной разрушительности с фундаментальной неопределенностью необходим содержательный и всеобъемлющий разговор о будущем ИИ и его влиянии на наш образ жизни. Надеюсь, моя книга внесет в него свой вклад.

Краткий словарь терминов

В нескольких интервью углубленно рассматриваются методы, используемые в сфере ИИ. Для понимания материала специальных знаний не требуется, но встречающиеся термины желательно знать. Вот объяснение наиболее важных из них. Если вы сочтете какой-то раздел технически сложным и запутанным, просто пропустите его и переходите к следующему.

Машинное обучение (**machine learning**) – раздел ИИ о методах построения алгоритмов, способных обучаться на данных. Другими словами, алгоритмы машинного обучения – это компьютерные программы, которые, по сути, программируют сами себя, просматривая информацию. Раньше считалось, что «компьютеры совершают только те действия, которые были запрограммированы», но эта ситуация меняется. Среди многочисленных типов алгоритмов машинного обучения самый революционный (и привлекающий всеобщее внимание) – это глубокое обучение.

Глубокое обучение (**deep learning**) – вид машинного обучения, в котором используются глубокие (или многоуровневые) **искусственные нейронные сети** (**artificial neural networks**), то есть программное обеспечение, имитирующее работу нейронов мозга. Глубокое обучение послужило основной движущей силой развития ИИ.

Есть и другие термины, которые, скорее всего, новичкам покажутся сложными. Без их глубокого понимания вполне можно обойтись, но краткое пояснение лишним не будет.

Метод обратного распространения ошибки (**backpropagation**) – алгоритм, используемый в системах глубокого обучения. Информация, поступающая в нейронную сеть, распространяется обратно через слои нейронов, вызывая у некоторых из них изменение настроек – весов (см. ниже «Обучение с учителем»). Так постепенно сеть находит правильный ответ. В 1986 г. Джеффри Хинтон стал соавтором первой полноценной статьи на эту тему, о чем более подробно вы узнаете из интервью с ним.

Еще более непонятный термин – **градиентный спуск** (**gradient descent**) – относится к математической технике, которую алгоритм обратного распространения использует для уменьшения ошибки в процессе обучения сети.

Встречаются в книге и термины, относящиеся к типам или конфигурациям нейронных сетей: **рекуррентные** (**recurrent**) и **сверточные** (**convolutional**) **сети**, а также **машины Больцмана** (**Boltzmann machines**). Различия обычно сводятся к способам связи нейронов. Детальное рассмотрение этих понятий выходит за рамки книги. Тем не менее я попросил объяснить их Яна Лекуна – изобретателя сверточной архитектуры, которая широко используется в распознавании объектов на изображениях.

Термин **байесовский** (**bayesian**) можно перевести как «вероятностный». Он встречается в таких сочетаниях, как «байесовские методы машинного обучения» или «байесовские сети». Они относятся к алгоритмам, которые используют вероятностные зависимости. Термин назван в честь священника Томаса Байеса (1701–1761), который сформулировал способ обновления вероятности события после возникновения другого, статистически взаимозависимого с ним. Байесовские методы очень популярны как среди специалистов по теории вычислительных машин и систем, так и среди ученых, моделирующих человеческое познание. Больше всего по этой теме рассказал Джуда Перл.

Способы обучения ИИ-систем

Существуют разные типы машинного обучения. Решающую роль в развитии искусственного интеллекта играют инновации, то есть новые способы обучения систем ИИ.

При **обучении с учителем** (**supervised learning**) алгоритму передаются структурированные, классифицированные и снабженные метками данные. Например, чтобы научить систему глубокого обучения распознавать на снимках собак, ей нужно предоставить много тысяч (или даже миллионов) изображений этого животного с меткой «собака». Кроме того, потребуется огромное количество изображений без собаки с меткой «нет собаки». После обучения можно показывать системе новые фотографии, и она будет определять наличие на них собаки на уровне, превосходящем возможности обычного человека.

Обучение с учителем – наиболее распространенный метод, применяемый в современных системах ИИ. На его долю приходится около 95 % практических приложений. Именно оно послужило основой машинного перевода (после обучения на миллионах предварительно переведенных документов) и ИИ-систем диагностики (после обучения на снимках с пометками «рак» и «нет рака»). К сожалению, для такого обучения требуются огромные объемы маркированных данных. Именно поэтому лидирующее положение в технологии глубокого обучения занимают такие компании, как Google, Amazon и Facebook.

Обучение с подкреплением (**reinforcement learning**), по сути, представляет собой обучение на практике или методом проб и ошибок. Система учится не на правильных размеченных данных, а самостоятельно ищет решение, получая подкрепление в случае успеха. Это напоминает дрессировку животных, которым в случае правильных действий дается кусочек вкусной еды. Именно обучение с подкреплением применялось для построения систем ИИ, играющих в игры. Из интервью с Демисом Хассабисом вы узнаете, что компания DeepMind использовала этот тип обучения для разработки компьютерной системы AlphaGo.

Проблема обучения по этому алгоритму заключается в необходимости огромного количества тренировочных попыток. Поэтому он применяется в основном для игр или для задач, которые можно воспроизводить на компьютере с высокой скоростью. Обучение с подкреплением можно использовать при разработке беспилотных автомобилей, но не для их эксплуатации на реальных дорогах. Виртуальные машины обучаются в искусственной среде, а после завершения обучения программное обеспечение устанавливается на реальные автомобили.

Обучение без учителя (**unsupervised learning**) обеспечивает непосредственное обучение на поступающих из окружающей среды неструктурированных данных. Именно так учатся люди. Например, дети учатся говорить, слушая речь родителей. Разумеется, человек использует и другие типы обучения, но самым характерным для него остается наблюдение и неконтролируемое взаимодействие с окружающей средой.

Обучение без учителя – один из наиболее многообещающих путей развития ИИ. Только представьте системы, умеющие обучаться сами без подготовки данных. Но их разработка – одна из самых сложных задач. Ее решение станет важной точкой на пути к созданию сильного ИИ.

Термин **сильный ИИ** обозначает истинно мыслящую машину, изначальную цель создания ИИ. Еще его называют интеллектом, сравнимым с человеческим разумом. Примеры сильного ИИ можно наблюдать в научной фантастике: компьютер HAL 9000 из «Космической одиссеи», главный компьютер космического корабля «Энтерпрайз» (или Дэйта) из «Звездного пути», андроид C3PO из «Звездных войн» и агент Смит из «Матрицы». Все эти вымышленные системы могли пройти **тест Тьюринга** (**Turing test**), то есть вести беседу как человек. Этот

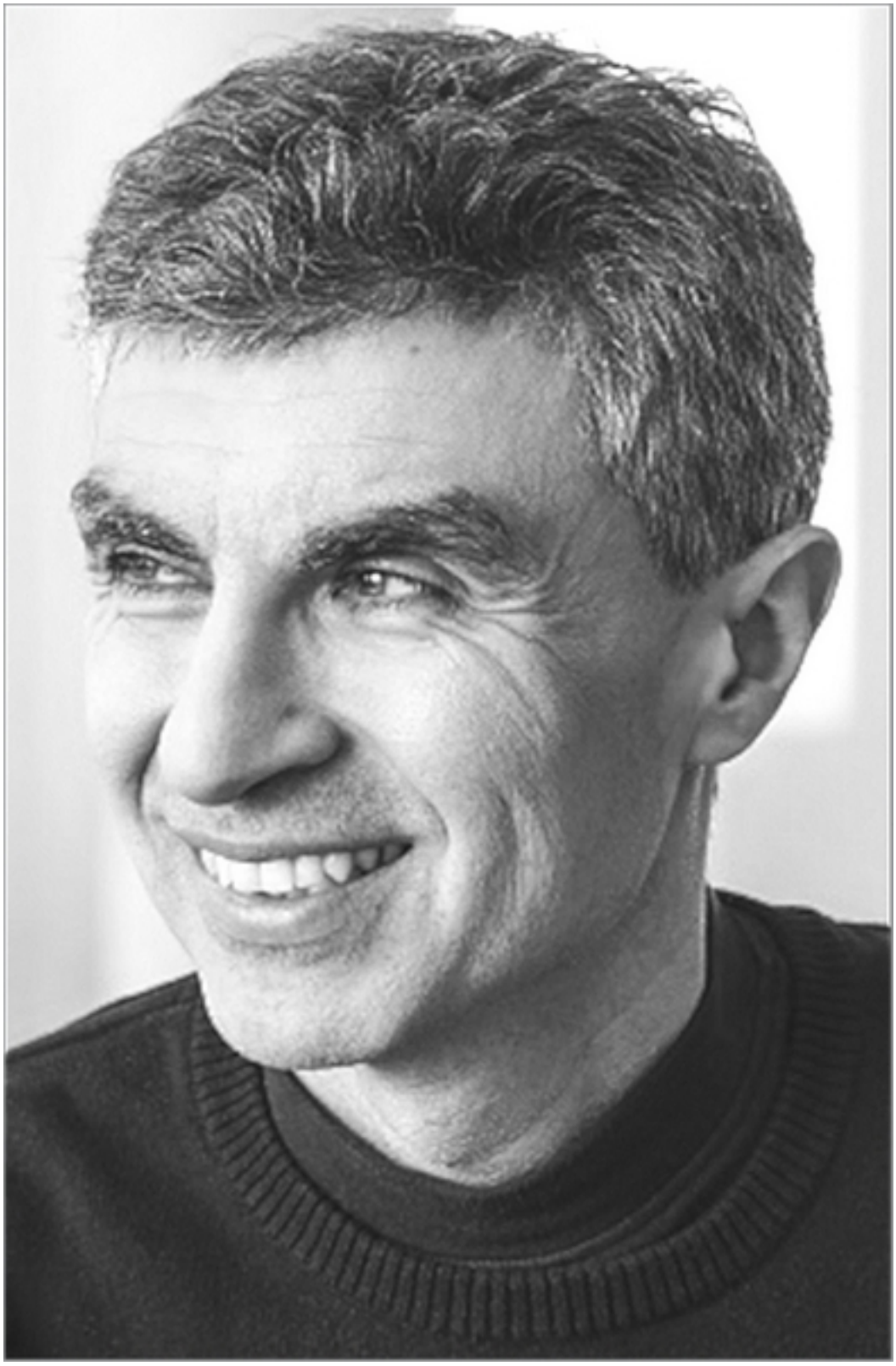
тест был предложен Аланом Тьюрингом в статье 1950 г. «Вычислительные машины и разум»[7], которую можно считать основополагающей работой в области ИИ.

Есть вероятность, что когда-нибудь появится **суперинтеллект** (**superintelligence**), или машина, превосходящая интеллектуальные способности любого человека. Это может произойти в результате простого увеличения аппаратных мощностей и быть ускорено самосовершенствованием этой машины. Так она запустит «рекурсивный цикл улучшения» или «быстрый интеллектуальный взлет», создавая проблему «выравнивания», если вступит в противоречие с интересам человека.

[7] Тьюринг А. Вычислительные машины и разум / Пер. с англ. К. Королева. – М.: АСТ, 2018. – 128 с. – (Серия «Эксклюзивная классика»).

Иошуа Бенджио

“ИИ, который существует сейчас и может появиться в обозримом будущем, не понимает и не чувствует нормы морали".



Директор Монреальского института алгоритмов обучения (MILA), доктор computer science, профессор кафедры информатики и математических методов Монреальского университета, соруководитель проекта Learning in Machines & Brains Канадского института перспективных исследований (CIFAR)

Иошуа Бенджио широко известен как один из пионеров глубокого обучения. Он активно продвигал исследования нейронных сетей, в частности обучение без учителя, и стал соавтором книги «Глубокое обучение»[8] *– одним из основных учебников по одноименному предмету.*

Мартин Форд: Вы играете ведущую роль в исследованиях ИИ, поэтому начать мне хотелось бы с вопроса о том, какие исследовательские проблемы стоят на пути к сильному ИИ.

Иошуа Бенджио: До создания ИИ, сравнимого с человеческим, нам еще очень далеко. Нужно понять, к примеру, почему невозможно создать машину, которая понимала бы окружающую действительность так же, как человек. Чего нам не хватает: обучающих данных или вычислительных мощностей? Многие считают, что причина состоит в отсутствии необходимых базовых компонентов, например, умения видеть причинно-следственные связи в данных, которое позволяет делать обобщения и находить правильные ответы в условиях, отличных от тренировочных.

Человек может представить, как он переживет новый для себя опыт. Например, если вы никогда не попадали в автомобильную аварию, вы все равно сможете прокрутить у себя в голове такую ситуацию и принять правильное решение. Обучение с учителем помогает компьютеру находить статистические закономерности в поставляемых данных, которые заранее классифицированы и размечены людьми.

Многие исследования пока не дали значимых результатов. Компьютер не может автономно приобретать знания о мире, воздействовать на него и наблюдать результат воздействия. Ответы на вопрос, как это реализовать, ищем не только мы.

М. Ф.: Какие проекты в настоящее время можно считать первостепенными в области глубокого обучения? Мне первым делом вспоминается программа AlphaZero. Есть ли другие?

И. Б.: На мой взгляд, из множества интересных проектов наиболее перспективны те, в которых агент в виртуальном мире пытается решать задачи, попутно изучая все с ними связанное. Такими проектами занимаемся мы в MILA, а также компании DeepMind, OpenAI, Университет Беркли, Facebook и Google в рамках проекта Google Brain. Это новые горизонты.

Но это долговременные исследования. Мы работаем не над конкретными вариантами применения глубокого обучения, а над тем, как научить агента осмысливать окружающую среду, говорить и понимать так называемый обоснованный язык (grounded language).

М. Ф.: Что означает этот термин?

И. Б.: Раньше компьютеры обучались языку, знакомясь с множеством текстов. Причем они достигали понимания только через связь слова с называемой им реалией. В отличие от робота, человек может сопоставить слово не только с объектом из реального мира, но и с вариантами изображения этого объекта.

Многочисленные исследования в области обучения обоснованному языку сводятся к попыткам научить роботов понимать язык хотя бы на уровне отдельных слов и выражений и реагировать соответствующим образом. Это очень интересное направление, необходимое для реализации таких вещей, как диалог с роботами, личные помощники и т. п.

М. Ф.: То есть, по сути, идея состоит в том, чтобы дать агенту свободу в смоделированной среде, позволив ему учиться, как это делают дети?

---

[8] *Бенджио И., Гудфеллоу Я., Курвилль А*. Глубокое обучение / Пер. с англ. А. Слинкина. – М.: ДМК пресс, 2017. – 652 с.: ил. Книга бесплатно доступна по адресу https://www.deeplearningbook.org.

И. Б.: Именно так. Более того, мы пользуемся результатами исследований в области детского развития и изучаем, какие этапы проходит новорожденный в первые месяцы жизни, постепенно приобретая представления о мире. До сих пор не совсем понятно, какие умения являются врожденными, а какие получены путем изучения.

Несколько лет назад я предложил для машинного обучения практику, которая используется при дрессировке животных – обучение по плану (curriculum learning). Обучающие примеры в этом случае демонстрируются не произвольно, а в последовательности, целесообразной для обучения. Процесс начинается с простых концепций, которые после их освоения учеником можно использовать как «кирпичики» для объяснения более сложных понятий.

М. Ф.: Я бы хотел поговорить о работе над сильным ИИ. Очевидно, что важной составляющей этого процесса вы считаете обучение без учителя. Что еще необходимо сделать?

И. Б.: Мой друг Ян Лекун сравнивает этот процесс с подъемом на гору. Сначала все радуются, насколько высоко забрались, но по мере приближения к вершине встречается множество других гор. Сейчас при разработке сильного ИИ четко видна ограниченность используемых подходов. Пока мы искали способы обучения более глубоких сетей, взбираясь на первую гору, создаваемые системы исследовались очень узко – на том этапе было важно просто подняться на несколько шагов вверх.

Как только применяемые техники обучения дали первые удовлетворительные результаты – мы приблизились к вершине первой горы, – стали заметны ограничения. И это следующая гора, которую нужно будет покорять. Поэтому невозможно сказать, сколько еще открытий потребуется.

М. Ф.: А вы можете хотя бы примерно оценить количество гор? Или период времени, который потребуется на создание сильного ИИ? Просто поделитесь своими прогнозами.

И. Б.: Не вижу смысла говорить о сроках. Невозможно предсказать, когда именно будет открыта дверь, от которой у нас нет ключа. Могу только заверить, что в ближайшие годы никаких прорывов не будет.

М. Ф.: Считаете ли вы перспективными глубокое обучение и нейронные сети в целом?

И. Б.: Да, многолетний прогресс в области глубокого обучения и нейронных сетей означает, что открытые концепции будут активно использоваться и дальше. Возможно, именно они помогут понять, каким образом мозг животных и человека осваивает сложные понятия. Но этого недостаточно для создания сильного ИИ. В настоящее время мы видим ограниченность имеющихся систем и собираемся улучшать и развивать их.

М. Ф.: Я знаю, что Институт искусственного интеллекта Пола Аллена (AI2) работает над проектом Mosaic, в рамках которого компьютеру пытаются помочь обрести разум. Считаете ли вы, что это важная задача? Ведь, возможно, разум рождается в процессе обучения?

И. Б.: Я уверен, что он возникает именно как результат обучения. Разум не может появиться только потому, что кто-то положил вам в голову какие-то знания. По крайней мере, у людей так.

М. Ф.: Глубокое обучение – основной путь к созданию сильного ИИ или потребуются гибридные системы?

И. Б.: Изначально ИИ был условным понятием, ни о каком обучении речи не шло. В центре внимания была способность машины делать последовательные выводы и объединять фрагменты информации. А глубокое обучение нейронных сетей можно назвать познанием снизу вверх. Все начинается с восприятия, в котором мы закрепляем понимание мира машиной. Затем можно строить распределенные представления и фиксировать связи между множеством переменных.

Отношения между такими переменными мы с братом изучали в 1999 г., что дало толчок к появлению в области естественного языка таких подходов, как векторное представление слов или распределенные представления слов и предложений. В них слово описывается характером

активности в мозге или набором чисел. Слова со сходными значениями связываются со сходными числовыми комбинациями.

В настоящее время на базе этих подходов пытаются решать классические проблемы ИИ, связанные с умением рассуждать и понимать, программировать и планировать. «Строительные блоки», обнаруженные при изучении восприятия, сейчас пробуют распространять на когнитивные задачи более высокого уровня (психологи называют это действиями Системы 2). Я полагаю, именно таким способом мы будем двигаться к сильному ИИ. Это нельзя назвать гибридной системой; скорее, мы пытаемся работать над классическим ИИ, используя как строительный материал концепции из глубокого обучения. Можно сказать, что требуются альтернативные пути достижения цели.

М. Ф.: То есть вы считаете, что все сведется к нейронным сетям с различными архитектурами?

И. Б.: Да. Ведь человеческий мозг состоит из нейронных сетей. Нужно придумать архитектуры и обучающие техники, позволяющие решать задачи, поставленные перед классическим ИИ.

М. Ф.: Обучения и тренировки будет достаточно или потребуется какая-то дополнительная структура?

И. Б.: Она уже существует, просто отличается от привычной структуры представления знаний, которую мы наблюдаем в энциклопедиях или формулах. Она имеет архитектуру нейронной сети и довольно широкие допущения по поводу окружающего мира и вершины собственных возможностей. Чтобы реализовывать в нейронной сети механизм внимания, такая структура требует большого количества предварительных знаний. Оказывается, данные имеют решающее значение для таких вещей, как машинный перевод.

Уже существует множество предположений в разных предметных областях о мире и о внедряемой функции, которые в виде архитектур и целей содержались в технологии глубокого обучения. Именно этому посвящено большинство современных научных работ.

М. Ф.: Говорят, что новорожденные развивают навык распознавания лиц с первых дней жизни. Очевидно, что это возможно благодаря некой структуре в мозге. Это не просто реакция нейронов на пикселы.

И. Б.: Ошибаетесь! Это именно реакция нейронов на пикселы, кроме того, в мозге ребенка присутствует особая структура, которая распознает нечто круглое с двумя точками внутри.

М. Ф.: Я считаю, что она существует с момента рождения.

И. Б.: Разумеется. И все то, что мы проектируем в нейронных сетях, тоже существует с самого начала. Работа исследователя в области глубокого обучения напоминает процесс эволюции. Знания вводятся как в виде структуры, так и через обучение.

При желании можно создать нечто, позволяющее сети распознавать лица, но в этом нет смысла, так как ИИ быстро обучается. Поэтому мы работаем над решением более сложных проблем.

Никто не говорит об отсутствии врожденных знаний у людей, детей и животных. Более того, у большинства животных знания исключительно врожденные. Муравью не приходится долго учиться, он действует в соответствии с заложенной в него программой. Но чем выше существо в иерархии интеллекта, тем большую роль в его жизнедеятельности начинает играть обучение. Человека отличает именно соотношение врожденных и приобретенных навыков.

М. Ф.: Я бы хотел уточнить некоторые из этих концепций. В 1980-е гг., после периода забвения, снова появился интерес к нейронным сетям, но речи о множестве слоев и глубине еще не шло. Вы участвовали в развитии глубокого обучения. Не могли бы вы простыми словами объяснить, что это такое?

И. Б.: Глубокое обучение – это совокупность методов машинного обучения. Но если в случае классического машинного обучения компьютеры учатся по прецедентам, глубокое обучение больше напоминает процесс, происходящий в мозге человека.

Эти методы работы над ИИ появились как продолжение более раннего изучения нейронных сетей. Слово «глубокие» указывает на появление у сетей дополнительных уровней, со своими вариантами представления информации. Мы надеемся, что углубление сетей позволит машине представлять более абстрактные вещи.

М. Ф.: То есть под слоями вы подразумеваете уровни абстракции? И если в качестве примера взять изображение, то первым уровнем будут пикселы, затем контуры и т. д.?

И. Б.: Да, все правильно.

М. Ф.: Правда ли то, что компьютеры до сих пор не понимают, что такое объект?

И. Б.: До некоторой степени компьютер понимает. Скажем, кошка понимает, что такое дверь, но не так, как человек. Даже люди обладают разными уровнями понимания многих вещей, а наука призвана углубить это понимание. Люди интерпретируют образы в контексте трехмерного мира благодаря стереоскопическому зрению и опыту познания. Человек получает не визуальную, а физическую модель объекта. Компьютер интерпретирует изображения на примитивном уровне, но для множества приложений этого достаточно.

М. Ф.: Правда ли, что глубокое обучение стало возможным благодаря методу обратного распространения ошибки, основная идея которого состоит в том, что информацию об ошибке можно отправить от выходов сети к ее входам, корректируя каждый слой в зависимости от конечного результата?

И. Б.: Да, метод обратного распространения стал краеугольным камнем успехов глубокого обучения. Он позволяет присваивать данным коэффициенты доверия (credit assignment), то есть рассчитывать, как для корректного поведения всей сети должны измениться внутренние нейроны. В контексте нейронных сетей об этом методе заговорили в начале 1980-х гг., когда я только начинал работать самостоятельно. Одновременно с Яном Лекуном метод развивали Джеффри Хинтон и Дэвид Румельхарт (David Rumelhart). Идея не новая, но примерно до 2006 г. особых успехов в обучении глубоких сетей не наблюдалось. Сейчас мы имеем механизм внимания, память и способность не только классифицировать, но и генерировать изображения.

М. Ф.: Существуют ли аналоги обратного распространения в человеческом мозге?

И. Б.: Хороший вопрос. Дело в том, что нейронные сети не пытаются скопировать мозг, хотя и появились как попытка смоделировать некоторые происходящие в нем процессы. Мы полностью не понимаем, как работает мозг. Нейробиологи пока не соединили результаты своих наблюдений в общую картину. Возможно, наша работа сможет дать доступную для проверки гипотезу. Ведь метод обратного распространения до сих пор считался уделом компьютеров, но не человеческого мозга. Прекрасные результаты, которые он дает, заставляют подозревать, что, возможно, мозг умеет проделывать похожие штуки. Я участвую в исследованиях, которые могут дать ответ на этот вопрос.

М. Ф.: В период «зимы ИИ», когда общий интерес к нему угас, вы вместе с Джеффри Хинтоном и Яном Лекуном продолжали свои исследования. Как вам удалось добиться таких успехов, как сейчас?

И. Б.: К концу 1990-х гг. нейронные сети вышли из моды, и ими практически никто не занимался. Но моя интуиция говорила, что мы упускаем что-то важное. Ведь благодаря композиционной структуре они могли представить богатую информацию о данных, базируясь на множестве «строительных блоков» – нейронов и их слоев. Лично меня это привело к лингвистическим моделям, то есть к нейронным сетям, которые моделировали текст, используя векторные представления слов. Каждое слово в них связано с набором чисел, соответствующих различным атрибутам, которые изучаются машиной автономно. Тогда этот подход не получил

широкого распространения, но в настоящее время эти идеи используются почти во всем, что связано с моделированием языка на основе данных.

Обучать глубокие сети мы не умели, но проблему решил Джеффри Хинтон своей работой по быстрым алгоритмам обучения ограниченной машины Больцмана (restricted Boltzmann machine, RBM). В моей лаборатории велась работа над связанными с ней автокодировщиками, которые дали начало таким моделям, как генеративно-состязательные сети (generative adversarial networks). Благодаря им появилась возможность обучения более глубоких сетей.

М. Ф.: А что такое автокодировщик?

И. Б.: Это специальная архитектура, состоящая из двух частей: кодировщика и декодера. То, что кодировщик сжал – декодер восстанавливал, причем так, чтобы выход был максимально близок к оригиналу. Автокодировщики превращали входную необработанную информацию в более абстрактное представление, в котором проще было выделить семантический аспект. Затем декодер восстанавливал по этой высокоуровневой абстракции исходные данные. Это были первые работы по глубокому обучению.

Через несколько лет мы обнаружили, что для обучения глубоких сетей достаточно изменения нелинейности. Вместе с одним из моих студентов, который работал с нейробиологами, мы решили попробовать блоки линейной ректификации (rectified linear unit, ReLU). Это пример копирования работы человеческого мозга.

М. Ф.: И к каким результатам это привело?

И. Б.: Раньше для активации нейронных сетей использовали сигмоиду, но оказалось, что с функцией ReLU гораздо проще обучать глубокие сети с большим количеством уровней. Переход случился примерно в 2010 г. Появилась огромная база данных ImageNet, предназначенная для отработки и тестирования методов распознавания объектов на изображениях и машинного зрения. Чтобы заставить людей поверить в методы глубокого обучения, нужно было показать хорошие результаты на примере этой базы. Это смогла сделать группа Джеффри Хинтона, которая использовала в качестве основы работы Яна Лекуна, посвященные сверточным сетям. В 2012 г. эти новые архитектуры позволили значительно улучшить существующие методы. За пару лет на эти сети переключились все, кто занимался компьютерным зрением.

М. Ф.: То есть именно в этот момент началось настоящее глубокое обучение?

И. Б.: Нет, совокупность факторов, ускоривших глубокое обучение, целиком сложилась только к 2014 г.

М. Ф.: То есть к моменту, когда этим занялись не только университеты, но и такие компании, как Google, Facebook и Baidu?

И. Б.: Именно так. Процесс ускорения начался чуть раньше, примерно в 2010 г., благодаря таким компаниям, как Google, IBM и Microsoft, которые работали над нейронными сетями для распознавания речи. Эти нейронные сети к 2012 г. Google начала использовать на смартфонах Android. Тот факт, что одну и ту же технологию глубокого обучения смогли применить как для компьютерного зрения, так и для распознавания речи, оказался по-настоящему революционным. Это привлекло внимание к сфере ИИ.

М. Ф.: Удивляет ли вас тот факт, что нейронные сети, с которыми вы много лет назад начали работать, стали центральным элементом проектов в таких крупных компаниях, как Google и Facebook?

И. Б.: Конечно, изначально этого никто не ожидал. В области глубокого обучения был сделан ряд важных, удивительных открытий. Я уже упоминал, что распознавание речи появилось в 2010 г., а о компьютерном зрении стали говорить в 2012 г. Пару лет спустя начался прорыв в сфере машинного перевода, который в 2016 г. привел к появлению сервиса Google Translate. В этом же году началось активное развитие программы AlphaGo. Всего этого мы не ожидали. Помню, как в 2014 г. я просматривал результаты генерации подписей к изображе-

ниям и поражался тому, что компьютер смог это сделать. Если бы годом раньше меня спросили, реально ли подобное, я бы ответил «нет».

М. Ф.: Это действительно нечто потрясающее. Конечно, осечки иногда происходят, но в большинстве случаев мы имеем поразительно точный результат.

И. Б.: Осечки неизбежны! Системы пока не обучены на достаточном количестве данных, кроме того, требуется изрядно продвинуться в фундаментальных исследованиях, чтобы они действительно научились распознавать объекты на изображениях и понимать язык. Пока до этого далеко, но ведь даже современного уровня производительности мы изначально не ожидали.

М. Ф.: А как вы пришли к исследованиям в области ИИ?

И. Б.: В юности я активно читал научную фантастику. Подозреваю, что это могло на меня повлиять. Именно оттуда я узнал об ИИ и трех законах робототехники Азимова, и у меня появилось желание изучать физику и математику. А чуть позже мы с братом заинтересовались компьютерами. На сэкономленные деньги мы приобрели компьютер Apple IIe, а затем Atari 800. Программного обеспечения тогда было мало, поэтому мы научились писать программы на языке BASIC.

Я так увлекся программированием, что занялся изучением вычислительной техники, а затем получил ученую степень в области computer science. В 1985 г., во время обучения в магистратуре, я начал читать статьи о первых нейронных сетях, в том числе работы Джеффри Хинтона. Это было любовью с первого взгляда. Я сразу понял, что хочу работать именно в этой сфере.

М. Ф.: Какой совет вы могли бы дать тем, что хочет заниматься глубоким обучением?

И. Б.: Прыгайте в воду и начинайте плавать. Сейчас информация любого уровня доступна в виде учебников, видео и библиотек с открытым исходным кодом. В сети можно бесплатно прочитать книгу «Глубокое обучение», соавтором которой я являюсь. В ней много информации для новичков. Студенты старших курсов зачастую тренируются, читая научные работы и пытаясь самостоятельно воспроизвести описанные там результаты, затем стараются попасть в лаборатории, проводящие исследования такого рода. Сейчас самое благоприятное время для карьеры в сфере ИИ.

М. Ф.: Из ключевых фигур в сфере глубокого обучения вы единственный, кто занимается только наукой. Большинство по совместительству сотрудничает с различными компаниями. Почему вы выбрали этот путь?

И. Б.: Я всегда высоко ценил научное сообщество, свободу работать на общее благо, делая вещи, которые, как я считаю, могут сильно повлиять на происходящее. Мне нравится работать со студентами, как психологически, так и с точки зрения продуктивности исследований. Уйти работать в индустрию – значит лишиться многого из этих вещей.

Кроме того, я хочу остаться в Монреале, а переход в индустрию означает переезд в Калифорнию или Нью-Йорк. Однажды я подумал, что можно попробовать создать новую Кремниевую долину для ИИ. В результате появился MILA, где проводятся фундаментальные исследования, задающие темп работы над ИИ во всем Монреале. Мы сотрудничаем с научно-исследовательским центром Vector Institute в Торонто и компанией Amii в Эдмонтоне в рамках канадской стратегии по продвижению ИИ в науке и экономике с пользой для социума.

М. Ф.: Раз уж вы упомянули об экономике, хотелось бы поговорить о рисках в этой сфере. Я много писал о том, что ИИ может привести к новой промышленной революции и потере множества рабочих мест. Как вы относитесь к этой гипотезе? Не преувеличена ли в данном случае угроза?

И. Б.: Нет, она не преувеличена. Непонятно только, когда это произойдет – в ближайшее десятилетие или намного позже. И даже если завтра мы полностью прекратим фундаментальные исследования в области ИИ, те результаты, которых мы уже достигли, позволят кому-то

получить социальное и экономическое преимущество за счет простого создания новых товаров и услуг.

Уже собрано огромное количество данных, которые мы пока не используем. Например, в здравоохранении применяется лишь малая доля доступной информации. А ее становится все больше, потому что каждый день оцифровываются новые данные. Производители аппаратных средств совершенствуют процессоры для глубокого обучения, что без сомнения изменит наш мир.

Конечно, прогресс в этой сфере замедляют социальные факторы. Общество не может моментально измениться, даже если технология идет вперед семимильными шагами.

М. Ф.: Реально ли решить проблему безработицы введением безусловного базового дохода?

И. Б.: Я думаю, это может сработать, но сначала нужно избавиться от морального ограничения, согласно которому у неработающего человека дохода быть не должно. Мне такая точка зрения кажется ненормальной. Думаю, нужно ориентироваться на то, что лучше для экономики и для счастья людей. Имеет смысл провести эксперимент, чтобы попытаться найти ответ на эти вопросы.

А единого ответа не будет. Позаботиться о людях, которые в результате новой промышленной революции останутся не у дел, можно разными способами. Мой друг Ян Лекун сказал, что если бы в XIX в. можно было предвидеть последствия промышленной революции, возможно, люди смогли бы избежать множества страданий. Если бы еще тогда, а не в 1950-х мы создали систему социальной защиты, которая сейчас существует в большинстве западных стран, сотни миллионов людей жили бы намного лучше. А ведь для новой революции, скорее всего, потребуется гораздо меньше столетия, и потенциальные негативные последствия могут быть еще сильнее.

Мне кажется, думать об этом нужно уже сейчас. Искать варианты, позволяющие минимизировать нищету и оптимизировать глобальное благополучие. Думаю, выход есть, но мы вряд ли его найдем, если будем держаться за старые ошибки и религиозные убеждения.

М. Ф.: Если все произойдет скоро, это станет еще и политической проблемой.

И. Б.: Поэтому нужно быстро реагировать!

М. Ф.: Совершенно справедливо. А чем еще, кроме влияния на экономику, может грозить ИИ?

И. Б.: Лично я активно выступал против роботов-убийц.

М. Ф.: Я слышал, что вы подписали письмо корейскому университету, который, по слухам, собирался заниматься их разработкой.

И. Б.: Да, и это помогло. Корейский институт передовых технологий (KAIST) сообщил, что они не будут разрабатывать автономные военные системы.

Отдельно я хотел бы коснуться такого важного вопроса, как включение людей в цикл управления. ИИ, который существует сейчас и может появиться в обозримом будущем, не понимает и не чувствует нормы морали. Разумеется, критерии добра и зла в разных культурах могут отличаться, тем не менее для людей они крайне важны.

Это касается не только роботов-убийц, но и роботов вообще. Представьте себе работу судьи. Для решения таких сложных моральных вопросов нужно понимать человеческую психологию и иметь моральные ценности. Нельзя передавать право принятия судьбоносных решений бездушной машине. Нужны социальные нормы или законы, гарантирующие, что в обозримом будущем компьютеры не получат таких полномочий.

М. Ф.: Здесь я мог бы с вами поспорить. Ваш взгляд на людей и их суждения крайне идеалистический.

И. Б.: Возможно. Но лично я предпочту, чтобы меня судил несовершенный человек, а не машина, не понимающая, что она творит.

М. Ф.: Но представьте себе автономного робота-охранника, который начинает стрелять только в ответ, когда в него попадает пуля. Человеку это недоступно, и потенциально именно такое поведение может спасти другие жизни. Теоретически, если запрограммировать такого робота правильно, он будет избавлен от расовых предрассудков. И в результате он получит преимущество перед человеком. Вы согласны?

И. Б.: Допускаю, что когда-нибудь такое станет возможным. Но вопрос в понимании машиной контекста задачи. Об этом компьютеры не имеют ни малейшего представления.

М. Ф.: Какие еще угрозы может нести ИИ?

И. Б.: Пока это практически не обсуждается, но после происшествий в Facebook и в Cambridge Analytica проблема может выйти на первый план. Она касается рекламы. Применение ИИ с целью воздействия на людей несет опасность для демократии и морально недопустимо. Общество должно позаботиться о предотвращении подобных явлений.

Например, в Канаде запрещена реклама, направленная на детей. Считается, что манипулировать уязвимыми умами аморально. Разумеется, уязвимы не только дети, иначе реклама бы просто не работала.

Во-вторых, реклама негативно влияет на состояние рынка, потому что за счет нее крупные компании мешают своим малоизвестным конкурентам. Современные технологии на базе ИИ позволяют еще точнее доносить посыл до целевой аудитории. Страшно то, что так людей можно заставить ухудшать собственную жизнь. Я имею в виду, например, политическую рекламу. С инструментами, которые позволяют влиять на людей, следует быть очень осторожными.

М. Ф.: Как вы можете прокомментировать предупреждения Илона Маска и Стивена Хокинга о смертельной угрозе, которую несет суперинтеллект, и о рекурсивном улучшении? Стоит ли об этом беспокоиться в данный момент?

И. Б.: Лично меня эти вопросы не волнуют. Исходя из текущего состояния дел, эти сценарии попросту нереалистичны. Они не совместимы с тем путем, по которому сейчас создается ИИ. Через несколько десятилетий все может измениться, но в настоящий момент – это научная фантастика. По крайней мере, с моей точки зрения. Более того, эти страхи отвлекают от насущных проблем, над которыми мы могли бы работать.

Кроме роботов-убийц и политической рекламы, существует, к примеру, проблема системной предвзятости в данных, ведущая к усилению дискриминации. Правительство и бизнес могут повлиять на это. Поэтому вместо обсуждения рисков, которые могут появиться в долгосрочной перспективе, нужно уделять внимание актуальным угрозам.

М. Ф.: А что вы думаете о работе в этой сфере, которую проводит Китай и другие страны? Например, вы упоминали об ограничениях на автономное оружие, но проблема в том, что некоторые страны могут игнорировать соглашение. Есть ли в данном случае повод для беспокойства?

И. Б.: Как ученый я не считаю это проблемой. Чем больше исследователей во всем мире работает над какой-то темой, тем лучше. Если Китай много инвестирует в исследования в сфере ИИ, это прекрасно; в конце концов, пользоваться результатами мы будем вместе. Хотя меня и пугают мысли о том, что китайское правительство может использовать технологию в военных целях. Системы, умеющие распознавать лица и следить за людьми, позволяют за считаные годы построить общество «Большого Брата». Технически это вполне осуществимо и представляет большую опасность для демократии. Это то, о чем мы должны беспокоиться. Подобное возможно при автократии.

Что же касается гонки вооружений, не нужно смешивать роботов-убийц и применение ИИ в военных целях. Я не считаю, что следует полностью запретить ИИ в армии. Если ИИ будет использован для создания оружия, уничтожающего роботов-убийц, это хорошо. Амо-

рально создание таких роботов, а не применение ИИ военными. Ведь работать можно и над оборонительным оружием.

М. Ф.: То есть вы считаете, что нужен свод правил работы над автономным оружием?

И. Б.: Свод правил требуется везде. По крайней мере, в областях, где применение ИИ будет влиять на общество. Нужно разработать правильные социальные механизмы, которые смогут гарантировать, что ИИ не будет использован во вред.

М. Ф.: И вы думаете, правительство в состоянии заняться этим вопросом?

И. Б.: Доверять решение этого вопроса компаниям точно не следует, потому что их в основном заботит увеличение прибыли. Конечно, они будут пытаться сохранить популярность у пользователей и клиентов, но их действия не совсем прозрачны.

Я думаю, что основную роль тут должно сыграть правительство, точнее даже международное сообщество.

М. Ф.: Считаете ли вы, что выгоды от ИИ в целом перевешивают связанные с ним риски?

И. Б.: Выгоды смогут перевесить риски, если мы будем действовать мудро. Именно поэтому так важно принимать правильные решения. И не хочется, закрыв глаза, мчаться вперед; нужно видеть все подстерегающие нас опасности.

М. Ф.: Где, по вашему мнению, все это должно обсуждаться? В аналитических центрах и университетах? Или требуется политическая дискуссия как на национальном, так и на международном уровне?

И. Б.: Нужна именно политическая дискуссия. В частности, на встрече Большой семерки, куда меня пригласили, был поставлен вопрос: «Какой путь развития ИИ может оказать положительное влияние на экономику и позволит сохранить доверие людей?» Потому что общество обеспокоено. И устранить это беспокойство поможет только открытая дискуссия, в которой смогут участвовать все желающие. Потому что ИИ и связанные с ним проблемы должны быть понятны любому человеку.

Стюарт Рассел

“Как только сильный ИИ «выйдет из детского сада», он превзойдет людей во всех возможных областях и будет обладать куда большей базой знаний, чем любой человек".

Профессор электротехники и computer science. директор Center for Intelligent Systems при Калифорнийском университете в Беркли

Стюарт Рассел известен как один из ведущих разработчиков ИИ. Является соавтором учебника по ИИ «Искусственный интеллект. Современный подход»[9]*, который в настоящее время используется более чем в 1300 колледжах и университетах в 118 странах. Получил степень бакалавра физики в Уодхэм-колледже Оксфордского университета и докторскую степень в области computer science в Стэнфорде. Занимался исследованиями на различные темы, связанные с ИИ, такие как машинное обучение, представление знаний и компьютерное зрение. Имеет многочисленные награды, в том числе Международной объединенной конференции по ИИ (IJCAI). Является членом Американской ассоциации содействия развитию науки, Ассоциации по продвижению ИИ (AAAI) и Ассоциации вычислительной техники (ACM).*

Мартин Форд: Вы написали учебник по ИИ, поэтому мне было бы интересно услышать, как вы определяете некоторые ключевые термины. Что входит в понятие ИИ? Какие проблемы информатики относятся к нему? Как ИИ связан с машинным обучением?

Стюарт Рассел: Я дам вам, скажем так, стандартное определение, которое приведено в нашей книге и в настоящее время общепризнано: «сущность разумна настолько, насколько правильно она поступает». Это означает, что ее действия должны приводить к поставленным целям. Определение относится как к людям, так и к машинам. Если разложить идею правильного поведения на составляющие и исследовать, окажется, что система ИИ должна уметь постигать, видеть, распознавать речь и действовать.

Еще требуется умение видеть суть вещей. Невозможно успешно функционировать в мире, о котором вам ничего не известно. Понять, каким образом мы осознаем различные вещи, помогает такое научное направление, как представление знаний. В его рамках изучаются способы внутреннего хранения данных, с последующей их обработкой алгоритмами формирования рассуждений, такими как алгоритмы автоматического логического вывода и вероятностного вывода.

Машинное обучение всегда было частью науки об ИИ. По сути, это развитие корректного поведения на базе предшествующего опыта.

М. Ф.: Еще дайте, пожалуйста, определения нейронным сетям и глубокому обучению.

С. Р.: Одна из стандартных методик машинного обучения – это обучение с учителем. Системе ИИ дается набор примеров какого-то понятия, снабженных описаниями и метками. Представьте фотографию с подписью, которая указывает, что это изображение лодки, далматинца или чашки с вишнями. Цель обучения состоит в поиске параметра или гипотезы, которые позволят классифицировать изображения в целом. Так мы пытаемся научить ИИ предсказывать, как могут выглядеть другие изображения тех же объектов.

Гипотезу или параметр можно представить в виде нейронной сети – схемы, состоящей из набора слоев. Входом в нее могут быть значения пикселов на фотографиях далматинцев. В процессе их распространения по схеме на каждом уровне вычисляются новые значения. На выходе из нейронной сети мы получаем распознавание объекта. И мы надеемся, что если подать на вход изображение далматинца, то после прохождения значений всех пикселов через все слои нейронной сети индикатор далматинца будет иметь высокое значение, а индикатор чашки с вишнями низкое. В этом случае можно сказать, что нейронная сеть правильно распознала объект.

М. Ф.: А как заставить нейронную сеть распознавать объекты на изображениях?

С. Р.: Для этого и нужен процесс обучения. Его алгоритмы настраивают весовые коэффициенты всех связей таким образом, чтобы на примерах сеть запоминала правильные ответы.

[9] Рассел С., Норвиг П. Искусственный интеллект. Современный подход. 2-е изд. / Пер. с англ. К. Птицына. – М.; СПб.: Диалектика, 2019. – 1407 с.: ил.

При определенном везении сеть начинает распознавать объекты и на новых, не входящих в обучающий набор изображениях.

Глубокое обучение – это обучение многослойных нейронных сетей. Формально минимального требования к глубине сети не существует, но двух- или трехуровневые сети, как правило, не считаются глубокими. Некоторые сети могут насчитывать более тысячи слоев. В них преобразование, происходящее между входом и выходом, можно представить как композицию более простых преобразований, происходящих на отдельных уровнях. Предполагается, что наличие множества уровней облегчает поиск обобщающих параметров благодаря установлению весовых коэффициентов всех связей.

Мы только подходим к теоретическому пониманию того, в каких случаях и почему глубокое обучение дает верные результаты. По большому счету все происходящее до сих пор выглядит для нас как магия. Кажется, что изображения, звуковые сигналы и речь, подаваемые на вход глубокой сети, обладают каким-то свойством, помогающим вычленить из них нужный признак. Но пока не ясно, каким.

М. Ф.: Может сложиться впечатление, что ИИ – это синоним глубокого обучения. Это не так?

С. Р.: Приравнивать глубокое обучение к ИИ – ошибка, потому что умение отличать далматинцев от ваз с вишнями – это малая часть требований к эффективному ИИ. Программы AlphaGo и AlphaZero привлекли внимание СМИ к глубокому обучению, но на самом деле это гибрид классического ИИ, который использует метод поиска, с алгоритмом глубокого обучения, который оценивает каждую игровую позицию. Хотя умение отличать хорошую позицию от плохой в го ключевое, программа не смогла бы сыграть на уровне чемпиона мира только в результате глубокого обучения.

По такому же принципу работает система беспилотного автомобиля. На дороге то и дело возникают ситуации, разрешение которых должно происходить по классическим правилам, но в то же время нужно предугадывать возможную реакцию других участников движения, оценивать последствия.

Восприятие – это важный компонент ИИ, который вполне адекватно удается реализовать через глубокое обучение, но для создания системы ИИ требуется множество других способностей различного типа. Особенно это касается действий, растянутых во времени, таких как поездка в отпуск, или сложных – строительство завода. Такие виды деятельности невозможно организовать, имея только систему типа «черный ящик» с глубоким обучением. Иначе алгоритму глубокого обучения нужно будет продемонстрировать все способы, которые когда-либо применялись для строительства. Научится ли система после этого строить заводы? Нет. Во-первых, таких данных не существует, а если бы они и были – нет смысла строить заводы таким образом.

Для строительства нужны специальные знания. Умение планировать. Знание свойств материалов. Чтобы решать долгосрочные и сложные задачи, можно создать системы ИИ, но глубокое обучение тут не поможет.

М. Ф.: Есть ли достижения в сфере ИИ, которые можно считать прорывом?

С. Р.: Хороший вопрос. Дело в том, что многие достижения, о которых активно говорили в СМИ, это не концептуальный прорыв, а всего лишь демонстрация. Вспомните хотя бы победу суперкомпьютера Deep Blue над Каспаровым. По сути, речь шла о демонстрации алгоритмов, разработанных тридцатью годами ранее и постепенно совершенствовавшихся на более мощном оборудовании. Но прорыв заключался в особенностях шахматной программы. В ней интересны и способ прогнозирования, и альфа-бета-алгоритм, сокращающий объем поиска, и некоторые из методов проектирования функций оценки. В итоге, как это часто бывает, СМИ назвали прорывом то, что им не является.

Также и сегодня. Вспомните отчеты о восприятии и распознавании надиктованной речи, заголовки в газетах о точности понимания текста на уровне человека или еще точнее. Но все эти впечатляющие практические результаты – только демонстрация прорывов, произошедших в 1980–1990-х гг.

Сейчас к более старым достижениям прибавлены современные методы проектирования, огромные наборы данных, многоуровневые сети и новейшее оборудование. Есть интерес к ИИ. Но обсуждаются не прорывы.

М. Ф.: Можно ли считать примером прорывной технологии программу AlphaZero от DeepMind?

С. Р.: Это интересная программа. Но нет ничего удивительного в том, что программное обеспечение для игры го смогли использовать для игры в шахматы и сеги на уровне чемпионов мира.

Тот факт, что программа AlphaZero менее чем за сутки научилась играть на сверхчеловеческом уровне в три разные игры, используя одно и то же программное обеспечение, безусловно, вызывает волнение. Но это всего лишь подтверждает, что если вы четко понимаете класс задачи, особенно детерминированной, если есть два игрока, делающих ходы по очереди, а игра идет по известным правилам и за ней можно наблюдать, то решением может стать хорошо спроектированный класс алгоритмов ИИ, позволяющих обучать функции оценки и использовать классические методы управления поиском.

Конец ознакомительного фрагмента.