Содержание

Часть III. ЗНАНИЯ, РАССУЖДЕНИЯ И ПЛАНИРОВАНИЕ

ГЛАВА 1

Введение

В этой главе авторы предпринимают попытку объяснить, почему они рассматривают искусственный интеллект как предмет, в наибольшей степени заслуживающий изучения, а также определить, в чем именно он заключается, — это необходимо сделать до того, как можно будет отправиться дальше.

Мы называем себя Homo Sapiens — *человек разумный*, — потому что наш ▶**интеллект**, наши умственные способности столь для нас важны. На протяжении тысячелетий люди пытались понять, *как мы думаем и действуем*, т.е. разобраться в том, как наш мозг, всего лишь небольшая горсточка материи, может ощущать, понимать, предсказывать и манипулировать миром, который несравнимо больше в размерах и сложнее, чем он сам. Область ▶**искусственного интеллекта**, или ИИ, охватывает не только понимание всего того, о чем говорилось выше, но и создание интеллектуальных сущностей — машин, которые будут способны вычислять, как им действовать эффективно и безопасно в самых разнообразных, в том числе незнакомых им, ситуациях.

Регулярно проводимые исследования свидетельствуют о том, что область ИИ расценивается как одна из самых интересных и наиболее быстро развивающихся областей науки и техники. Уже сейчас она приносит годовой доход размером более триллиона долларов. Эксперт по искусственному интеллекту Кай-Фу Ли предсказывает, что ее влияние будет "больше, чем что-либо иное в истории человечества". Более того, интеллектуальные границы ИИ широко открыты. В то время как студенты, изучающие традиционные науки, такие как физика, могут полагать, что лучшие идеи в этой области уже были выдвинуты Галилеем, Ньютоном, Кюри, Эйнштейном и другими, они осознают, что в области ИИ еще достаточно простора для выдающихся открытий.

Тематика области искусственного интеллекта в настоящее время охватывает огромный перечень научных направлений, от задач самого общего характера (обучение, рассуждение, восприятие и т.д.) и до таких конкретных задач, как игра в шахматы, доказательство математических теорем, сочинение стихов, вождение автомобиля или диагностика заболеваний. Достижения в области ИИ могут найти себе применение при решении любой интеллектуальной задачи, — это универсальная научная область.

1.1. Что такое ИИ

Выше мы уже заявили, что область ИИ вызывает большой интерес, но пока еще не пояснили, что же она собой представляет. Исторически сложилось так, что исследователи рассматривали несколько различных версий ИИ. Одни давали определение интеллекту с точки зрения соответствия поведению человека, в то время как другие предпочитали использовать абстрактное, формальное определение интеллекта, получившее название ►**рациональность** — в широком смысле это способность "поступать правильно". Сам предмет также воспринимается по-разному: одни считают, что интеллект является свойством внутренних *мыслительных процессов* и *рассуждений*, в то время как другие фокусируются на интеллектуальном *поведении*, т.е. на внешней характеристике.[1]

Из этих двух противопоставлений — *человекоподобность–рациональность*[2] и *мышление–поведение* можно вывести четыре различные попарные комбинации, и у каждой из них будут свои приверженцы и соответствующие исследовательские программы. Используемые в них методы были по необходимости различными: поиски человекоподобного интеллекта должны были проводиться в рамках эмпирических наук, связанных с психологией, включая наблюдение и гипотезы о фактическом человеческом поведении и мыслительных процессах. С другой стороны, рационалистический подход предполагает некий синтез математики и техники, с привлечением статистики, теории управления и экономики. Группы исследователей, следовавшие различными путями, могли как проявлять пренебрежение, так и помогать друг другу. Давайте рассмотрим все четыре подхода более подробно.

1.1.1. Действуя, как человек: подход на основе теста Тьюринга

►**Тест Тьюринга**, предложенный Аланом Тьюрингом в 1950 году, был разработан как мысленный эксперимент, который позволил бы обойти философскую неясность вопроса "Может ли машина мыслить?" Компьютер пройдет этот тест, если человек-испытатель, направив ему несколько письменных вопросов, в конечном итоге не сможет определить, от кого исходят полученные им письменные ответы — от человека или от компьютера. В главе 27 подробно обсуждается этот тест и рассматривается вопрос о том, действительно ли можно считать интеллектуальным компьютер, который успешно прошел подобный тест. На данный

[1] Не следует смешивать понятия "искусственный интеллект" и "машинное обучение". Машинное обучение — это область ИИ, в которой изучается способность улучшать свои навыки на основе опыта. В одних системах искусственного интеллекта используются методы машинного обучения для достижения необходимого уровня знаний, а в других этот подход не используется.

[2] Мы не предполагаем, что люди "иррациональны" в буквальном смысле этого слова, т.е. "лишены нормальной ясности ума". Мы просто допускаем, что человеческие решения не всегда безупречны с точки зрения математики.

момент просто отметим, что программирование компьютера для прохождения этого теста в строгом соответствии с исходными требованиями потребует очень большого объема работы. Запрограммированный таким образом компьютер должен обладать всеми перечисленными ниже возможностями.

- ►**Обработка естественного языка** для успешного общения с человеком на его языке.
- ►**Представление знаний** для успешного сохранения того, что он узнает или услышит.
- ►**Автоматические рассуждения** для ответа на вопросы и вывода новых заключений.
- ►**Машинное обучение** для адаптации к новым обстоятельствам, а также для выявления и экстраполяции моделей.

Сам Тьюринг полагал, что для демонстрации искусственного интеллекта нет необходимости в физической имитации человека. Однако другие исследователи с этим не согласились и предложили ►**общий тест Тьюринга**, для прохождения которого необходимо продемонстрировать взаимодействие с объектами и людьми в реальном мире. Чтобы пройти полный тест Тьюринга, роботу дополнительно понадобятся следующие способности.

- ►**Компьютерное зрение** и распознавание речи для восприятия реального мира.
- ►**Робототехника** для манипулирования объектами и перемещения в пространстве.

Эти шесть перечисленных выше направлений составляют основную часть области исследований ИИ. Тем не менее исследователи искусственного интеллекта практически не занимаются решением задачи прохождения теста Тьюринга, считая, что гораздо важнее изучить основополагающие принципы интеллекта. И действительно, проблему “искусственного полета” удалось успешно решить лишь после того, как инженеры и изобретатели перестали имитировать птиц и приступили к изучению аэродинамики. В научных и технических работах по воздухоплаванию цель этой области знаний не определяется как “создание машин, которые в своем полете настолько напоминают голубей, что даже могут обмануть настоящих птиц этого вида”.

1.1.2. Думая, как человек: подход когнитивного моделирования

Чтобы сказать, что программа мыслит, как человек, мы должны знать, как люди думают. Мы можем изучать человеческое мышление тремя способами.

- ►**Самоанализ** — попытки поймать наши собственные мысли, когда они приходят в наше сознание.

- ►**Психологические эксперименты** — наблюдение за человеком в действии.
- ►**Визуализация работы мозга** — наблюдение за мозгом в действии.

Если у нас появится достаточно точная теория работы сознания, станет возможным выразить эту теорию в виде компьютерной программы. Если поведение системы ввода-вывода программы соответствует действительному поведению человека, это свидетельствует о том, что и некоторые механизмы программы могут работать, как у людей.

Например, Аллен Ньюэлл (Allen Newell) и Герберт Саймон (Herbert Simon), которые разработали программу GPS (General Problem Solver — универсальный решатель задач) ([1668], 1961), не стремились лишь к тому, чтобы эта программа правильно решала поставленные задачи. Они в большей степени заботились о том, чтобы запись этапов проводимых ею рассуждений совпадала с регистрацией рассуждений людей, решающих такие же задачи. Междисциплинарная область ►**когнитивной науки** объединяет компьютерные модели из ИИ и экспериментальные методы из психологии для построения точных, позволяющих выполнить их проверку теорий человеческого разума.

Когнитивная наука сама по себе является увлекательной областью, достойной нескольких учебников и по крайней мере одной энциклопедии ([2354], 1999). Время от времени мы будем комментировать сходства или различия между методами искусственного интеллекта и человеческим познанием. Однако реальная когнитивная наука по необходимости строится на основе экспериментальных исследований реальных людей или животных. Мы оставим обсуждение этого аспекта для других книг, поскольку предполагаем, что для проведения экспериментов читатель располагает только компьютером.

На ранних этапах исследований ИИ часто возникала путаница между разными подходами. Автор мог утверждать, что алгоритм хорошо справляется с заданием и, *следовательно*, является хорошей моделью человеческой деятельности, или наоборот. Современные авторы разделяют эти два вида претензий; это различение позволяет как ИИ, так и когнитивной науке развиваться более быстрыми темпами. Эти две области исследований часто оплодотворяют друг друга, что наиболее заметно в компьютерном зрении, где результаты нейрофизиологических исследований используются при построении вычислительных моделей. В последнее время комбинирование методов нейровизуализации с технологиями машинного обучения с целью анализа собираемых данных уже привело к появлению возможности “читать мысли”, т.е. к возможности определения семантического содержания мыслей в сознании человека. Эта способность, в свою очередь, могла бы пролить дополнительный свет на то, как работает человеческое познание.

1.1.3. Думая рационально: подход на основе "законов мышления"

Древнегреческий философ Аристотель был одним из первых, кто попытался определить законы "правильного мышления", т.е. процессы формирования неопровержимых рассуждений. Его ►**силлогизмы** стали образцом для создания процедур доказательства, которые всегда позволяют прийти к правильным заключениям, если даны правильные предпосылки. Вот канонический пример таких рассуждений: "Сократ — человек; все люди смертны; следовательно, Сократ смертен". (Этот пример, вероятно, скорее связан с Секстом Эмпириком, чем с Аристотелем.) Предполагалось, что эти законы мышления управляют работой ума; их исследование положило начало научному направлению, называемому **логикой**.

В XIX веке ученые-логики создали точную систему логических обозначений для утверждений о предметах любого рода, встречающихся в мире, и об отношениях между ними. (Сравните это с обычной системой арифметических обозначений, которая предназначена в основном для формирования утверждений о равенстве и неравенстве *чисел*.) К 1965 году уже были разработаны программы, которые в принципе могли решить любую разрешимую проблему, описанную в системе логических обозначений. Исследователи в области искусственного интеллекта, придерживающиеся так называемых традиций ►**логицизма**, надеются, что им удастся создать интеллектуальные системы на основе подобных программ.

Логика, как она обычно понимается, требует, чтобы знания о мире, которыми она оперирует, были *точными* — условие, которое в действительности редко достижимо. Мы просто не знаем правил, которые действуют, скажем, в политике или на войне, с той же степенью достоверности, как правила арифметики или игры в шахматы. ►**Теория вероятности** заполняет этот пробел, позволяя проводить строгие рассуждения с неточной информацией. В принципе, это позволяет построить всеобъемлющую модель рационального мышления, которая обеспечит переход от необработанной субъективно воспринимаемой информации к пониманию того, как устроен мир и даже к предсказаниям о будущем. Но этой модели все же недостаточно для генерирования разумного *поведения*. Для этого нам потребуется еще и теория рационального действия — рационального мышления самого по себе нам будет недостаточно.

1.1.4. Действуя рационально: подход с использованием рационального агента

►**Агент** — это просто что-то, что действует (слово *агент* произошло от латинского слова *agere* — "действовать"). Конечно, все компьютерные программы что-то делают, но ожидается, что компьютерные агенты будут делать больше: работать автономно, воспринимать окружающую среду, сохранять свое существование в течение длительного периода времени, приспосабливаться к изменениям,

устанавливать и преследовать определенные цели. ►**Рациональным агентом** называется агент, действующий таким образом, чтобы достичь наилучшего результата или, если он находится в условиях неопределенности, наилучшего ожидаемого результата.

В подходе к созданию ИИ на основе "законов мышления" акцент был сделан на формировании правильных логических выводов. Безусловно, иногда формирование правильных логических выводов становится *частью* функционирования и рационального агента, поскольку один из способов рациональной организации своих действий состоит в том, чтобы логическим путем прийти к заключению, что данное конкретное действие позволяет достичь указанных целей, а затем действовать в соответствии с принятым решением. С другой стороны, существуют способы действовать рационально, о которых нельзя сказать, что они предполагают логический вывод. Например, отдергивание пальца от горячей печи — это рефлекторное действие, которое обычно является более успешным в сравнении с более медленным действием, предпринятым после тщательного обдумывания ситуации.

Все навыки, необходимые для теста Тьюринга, также позволяют агенту действовать рационально. Представление знаний и рассуждения обеспечат агенту возможность принимать правильные решения. Мы должны быть в состоянии генерировать понятные фразы на естественном языке, чтобы войти в состав сложного социума. Нам нужно учиться не только для эрудиции, но и для развития способности генерировать эффективное поведение, особенно в новых условиях.

Подход к созданию ИИ с использованием рациональных агентов имеет два важных преимущества в сравнении с другими подходами. Во-первых, он более общий, чем подход на основе "законов мышления", — правильный вывод является лишь одним из нескольких возможных механизмов достижения рациональности. Во-вторых, он лучше поддается научному развитию. Стандарт рациональности в самом общем виде хорошо определен математически. Во многих случаях можно исходить из этой спецификации, чтобы получить проект агента, который доказуемо достигнет цели, что почти невозможно, если цель состоит в том, чтобы имитировать человеческое поведение или процессы мышления.

По этим причинам подход к созданию ИИ с использованием рациональных агентов преобладал на протяжении большей части истории проведения исследований в этой области. В первые десятилетия рациональные агенты строились на логических основах и формировали определенные планы для достижения конкретных целей. Позже методы, основанные на теории вероятностей и технологии машинного обучения, позволили создавать агентов, которые были способны принимать решения в условиях неопределенности, имея целью достижение наилучшего ожидаемого результата. Сказанное можно обобщить следующим образом: ➔ *большинство работ в области ИИ фокусировалось, прежде всего, на изучении и создании агентов, способных* ►***поступать правильно***. Что считать правильным, определялось целью, которая ставилась перед агентом. Эта общая парадигма настолько распространена, что ее вполне правомерно назвать ►**стандартной моделью**. Она превалирует не только в области

ИИ, но и в теории управления, где регулятор минимизирует стоимостную функцию, в исследовании операций, где линия поведения максимизирует целевую функцию, в статистике, где правило принятия решения минимизирует функцию потерь, и в экономике, где лицо, принимающее решения, максимизирует полезность или некоторую меру социального обеспечения.

В отношении стандартной модели необходимо сделать одно важное уточнение: следует учесть тот факт, что идеальная рациональность — всегда выбирать именно оптимальное действие — не всегда достижима в сложных условиях. Например, требования к вычислительным ресурсам могут оказаться слишком высокими. В главах 5 и 17 будет обсуждаться вопрос ▶**ограниченной рациональности** — как поступить надлежащим образом, если не хватает времени на проведение всех необходимых вычислений. Тем не менее идеальная рациональность часто остается хорошей отправной точкой для проведения теоретического анализа.

1.1.5. Полезные машины

Стандартная модель была полезным ориентиром для исследований в области ИИ с самого начала, но в долгосрочной перспективе она, вероятно, уже не будет настолько подходящей. Причина в том, что стандартная модель предполагает, что машине всегда ставится точно определенная цель.

Для искусственно определенных задач, таких как игра в шахматы или нахождение кратчайшего пути, задача формулируется с изначально определенной конкретной целью, поэтому стандартная модель здесь будет вполне применима. Однако по мере приближения к реальному миру становится все труднее и труднее определить конечную цель точно и полностью. Например, при проектировании самоуправляемого автомобиля изначально можно полагать, что цель состоит лишь в том, чтобы безопасно достичь пункта назначения. Однако движение по любой дороге сопряжено с риском получения травмы из-за других движущихся по ней автомобилей, отказа оборудования и т.д. В результате жестко заданная цель обеспечения полной безопасности приводит к единственному оптимальному решению: просто оставаться в гараже. Необходим некий компромисс между требованием достижения прогресса в отношении приближения к месту назначения и риском получить при этом травму. Как можно достичь такого компромисса? Пойдем дальше: в какой степени мы можем позволить самоуправляемой машине совершать действия, которые будут раздражать других водителей? В какой степени автомобиль должен смягчать ускорение, крутизну поворотов и резкость торможения, чтобы исключить неприятные ощущения у пассажира? На такие вопросы сложно ответить априори. Эти и другие подобные аспекты создают множество проблем во всей области взаимодействия человека и робота, где самоуправляемый автомобиль является лишь одним из примеров.

Проблема достижения согласия между нашими истинными предпочтениями и той целью, которую мы ставим перед машиной, называется ▶**проблемой**

выравнивания ценностей: ценности или цели, передаваемые машине, должны быть согласованы с ценностями человека. Если система искусственного интеллекта разрабатывается в лаборатории или в симуляторе — как это и было раньше в большинстве исследований в данной области, — будет очень просто исправить неверно выбранную цель: сбросить систему, откорректировать цель и попробовать еще раз. Но по мере того, как исследования в этой области проводятся со все более и более сложными и умными интеллектуальными системами, развертываемыми в реальном мире, такой подход становится нежизнеспособным. Развертывание системы с неверно заданной целью неизбежно будет иметь негативные последствия. Более того, чем более интеллектуальной будет такая система, тем более отрицательными будут последствия.

Возвращаясь к явно беспроблемному примеру игры в шахматы, рассмотрим, что произойдет, если машина будет достаточно умна, чтобы рассуждать и действовать и за пределами шахматной доски. В этом случае она может попытаться увеличить свои шансы на победу, используя такие хитрости, как использование гипноза или шантаж своего оппонента, либо даже подкуп аудитории, чтобы она шумела в то время, когда противник будет размышлять над очередным ходом.[3] Она даже может попытаться захватить для себя дополнительные вычислительные мощности. ➔ *Такое поведение не является "неразумным" или "безумным", в действительности оно является логическим следствием определения победы как единственной цели машины.*

Невозможно противодействовать всем способам неправильного поведения машины, преследующей фиксированную цель. И это весомая причина, чтобы прийти к заключению, что стандартная модель является неадекватной. Мы не хотим машин, которые будут интеллектуальными в смысле преследования *их* целей; мы хотим, чтобы они преследовали *наши* цели. Если мы не можем точно передать эти цели машине, то нам нужна новая формулировка — такая, согласно которой машина преследует наши цели, но *обязательно* не имеет полной уверенности в том, каковы они. Когда машина знает, что ей не известны цели во всей их полноте, у нее будет стимул действовать осторожно, просить разрешения на те или иные действия, чтобы узнать больше о наших предпочтениях посредством наблюдения, и считаться с контролем со стороны человека. В конечном счете мы хотим агентов, которые будут ►**доказуемо полезны** человеку. Мы вернемся к этой теме в разделе 1.5.

1.2 Истоки искусственного интеллекта

В этом разделе кратко описана история развития научных дисциплин, которые внесли свой вклад в область ИИ в виде конкретных идей, воззрений и методов. Как и в любом историческом очерке, поневоле приходится ограничиваться описанием небольшого круга людей, событий и открытий, игнорируя все остальные

[3] В одной из первых книг по шахматам Руи Лопес (1561) писал: "Всегда ставьте доску так, чтобы солнце светило в глаза вашему противнику".

факты, которые также были важны. Авторы построили этот исторический экскурс вокруг ограниченного круга вопросов. Безусловно, они не хотели бы, чтобы у читателя создалось такое впечатление, будто эти вопросы являются единственными, которые рассматриваются в указанных научных дисциплинах, или что сами эти дисциплины развивались исключительно ради того, чтобы их конечным итогом стало создание искусственного интеллекта.

1.2.1. Философия

- Можно ли использовать формальные правила для получения обоснованных заключений?
- Как мысль возникает в физическом мозге?
- Откуда приходят знания?
- Каким образом знание ведет к действию?

Аристотель (384–322 до н.э.) был первым, кто сформулировал точный свод законов, регулирующих рациональную часть нашего мышления. Он разработал неформальную систему силлогизмов, предназначенную для проведения правильных рассуждений, которая в принципе позволяла любому делать выводы механически, лишь на основании начальных предпосылок.

Раймон Луллий (ок. 1232–1315) разработал систему рассуждений, опубликованную им под названием *Ars Magna,* или *Великое искусство* ([1438], 1305). Луллий даже предпринял попытку реализовать свою систему в виде механического устройства: набора бумажных колец, которые можно было поворачивать, получая разные перестановки.

Около 1500 года Леонардо да Винчи (1452–1519) спроектировал, но не построил механический калькулятор. Недавние реконструкции этого устройства показали, что оно вполне работоспособно. Первая известная машина для выполнения расчетов была создана примерно в 1623 году немецким ученым Вильгельмом Шиккардом (1592–1635). В 1642 году Блез Паскаль (1623–1662) построил арифметическую машину “Паскалин”. Он писал, что она “производит эффект, который кажется более близким к мышлению по сравнению с любыми действиями животных”. Готфрид Вильгельм Лейбниц (1646–1716) создал механическое устройство, предназначенное для выполнения операций над *концепциями*, а не числами, но область его применения была довольно ограниченной. В своей книге *Левиафан*, вышедшей в 1651 году, Томас Гоббс (1588–1679) выдвинул идею создания думающей машины, “искусственного животного”, как он ее называл. По его словам, “Вместо сердца у нее будет пружина, вместо нервов — пучок струн, а вместо суставов — множество колес”. Он также предположил, что рассуждение подобно числовым расчетам: “Ведь «суждение»… это не что иное, как «подведение итогов», в ходе которого мы складываем и вычитаем”.

Одно дело — сказать, что сознание функционирует, по крайней мере частично, в соответствии с логическими или числовыми правилами, а затем построить

физические системы, которые имитируют некоторые из этих правил. Совсем другое дело — сказать, что сознание само по себе является такой физической системой. Рене Декарт (1596–1650) впервые опубликовал строгое обсуждение различий между разумом и материей. Он отметил, что чисто физическая концепция ума, похоже, оставляет мало места для свободной воли. Если сознание регулируется исключительно физическими законами, то оно имеет не больше свободной воли, чем скала, "решившая" рухнуть вниз. Декарт был сторонником ▶**дуализма**. Он считал, что есть часть человеческого сознания (или *душа* либо *дух*), которая находится за пределами естества и не подчиняется физическим законам. Животные, с другой стороны, не обладают таким дуалистическим свойством, поэтому их можно рассматривать как машины.

Альтернативой дуализму является **материализм**, утверждающий, что сознание *складывается* из операций, выполняемых мозгом в соответствии с законами физики. Свободная воля — это просто форма, которую принимает восприятие нашим существом доступных вариантов в процессе выбора. Для описания подобного представления, исключающего любую возможность существования сверхъестественного, также используются термины **физикализм** и **натурализм**.

Если полагать, что знаниями манипулирует физический разум, то возникает следующая проблема — установить источник знаний. Такое научное направление, как ▶**эмпиризм**, родоначальником которого был Френсис Бекон (1561–1626), автор *Нового Органона*,[4] можно охарактеризовать высказыванием Джона Локка (1632–1704): "В человеческом понимании нет ничего, что не проявлялось бы прежде всего в ощущениях".

Дэвид Юм (1711–1776) в своей книге *Трактат о человеческой природе* ([1095], 1739) предложил метод, известный теперь под названием ▶**принцип индукции**, — общие правила вырабатываются путем изучения повторяющихся ассоциаций между элементами, которые рассматриваются в этих правилах.

Основываясь на работе Людвига Витгенштейна (1889–1951) и Бертрана Рассела (1872–1970), знаменитый Венский кружок, группа философов и математиков, собиравшихся в Вене в 1920–1930-е годы, разработал доктрину ▶**логического позитивизма**. Согласно этой доктрине все знания могут быть охарактеризованы с помощью логических теорий, связанных в конечном итоге с ▶**протокольными предложениями**, которые соответствуют наблюдаемым фактам. Таким образом, логический позитивизм объединяет рационализм и эмпиризм.

В ▶**теории подтверждения** Рудольфа Карнапа (1891–1970) и Карла Хемпеля (1905–1997) была предпринята попытка понять, как знания могут быть приобретены из опыта посредством количественной оценки степени доверия, присваиваемой логическим предложениям на основе сопоставления с наблюдениями, подтверждающими или опровергающими их. В книге Карнапа *Логическая структура*

[4] Книга *Новый органон* была создана как новая версия труда Аристотеля *Органон* (инструмент мышления).

мира ([372], 1928) была сформулирована, по-видимому, первая теория мышления как вычислительного процесса.

Последним элементом в философской картине разума является связь между знанием и действием. Этот вопрос является жизненно важным для искусственного интеллекта, поскольку интеллектуальность требует не только рассуждений, но и действий. Более того, только понимая, как обосновать предпринимаемые действия, можно понять, как создать агент, действия которого будут обоснованы (или рациональны).

Аристотель утверждал (в *De Motu Animalium*), что действия обосновываются логической связью между целями и знанием о результатах этих действий.

> Но почему происходит так, что размышления иногда сопровождаются действием, а иногда — нет, иногда за ними следует движение, а иногда — нет? Создается впечатление, что почти то же самое происходит и в случае построения рассуждений и формирования выводов о неизменных объектах. Но в таком случае целью умственной деятельности оказывается умозрительное суждение… тогда как заключением, которое следует из данных двух предпосылок, является действие… Мне нужна защита от дождя; защитой может послужить плащ. Мне нужен плащ. Я должен изготовить то, в чем нуждаюсь; я нуждаюсь в плаще. Я должен изготовить плащ. И заключение "я должен изготовить плащ" становится действием.

В книге *Никомахова этика* (том III. 3, 1112b) Аристотель дополнительно развивает эту тему, предлагая следующий алгоритм.

> Мы размышляем не о конечных целях, а о средствах. Врач не обдумывает, должен ли он лечить, а оратор — должен ли он убедить... Они уже установили конечную цель и рассматривают, как и за счет чего она достигается, и если окажется несколько средств, то определяют, какое из них самое простое и наилучшее; если же достижению цели служит одно средство, думают, *как* ее достичь при помощи этого средства и что будет средством для *этого* средства, пока не дойдут до первой причины, которую находят последней… и то, что является последним в порядке анализа, окажется первым в порядке выполнения. А если мы сталкиваемся с невозможностью, то прекращаем поиск — например, если нам нужны деньги, а их нельзя получить, — но если что-то кажется возможным, мы пытаемся это сделать.

Алгоритм Аристотеля был реализован через 2300 лет Ньюэллом и Саймоном в их программе **General Problem Solver** (GPS). Теперь то, что создано на его базе, принято называть системой жадного регрессивного планирования (см. главу 11). Методы, основанные на логическом планировании для достижения определенных целей, доминировали в первые несколько десятилетий теоретических исследований в области ИИ.

Анализ исключительно с точки зрения действий по достижению цели часто является полезным, но иногда оказывается неприменимым. Например, если к цели ведет несколько вариантов действий, необходимо иметь какой-то способ выбирать среди них. Еще важнее то, что иногда может не быть полной уверенности в возможности достижения цели, но некоторые действия все же следовало бы

предпринять. Как в таких ситуациях следует поступать? Антуан Арно ([74], 1662), анализируя идею принятия рациональных решений в азартных играх, предложил количественную формулу максимизации ожидаемого конечного денежного результата. Позже Даниэль Бернулли ([191], 1738) ввел более общее понятие ▶**полезности** для фиксации внутренней, субъективной ценности результата. Современное понятие рационального принятия решений в условиях неопределенности предполагает максимизацию ожидаемой полезности, как это описывается в главе 16.

В вопросах этики и государственной политики лицо, принимающее решения, должно учитывать интересы множества людей. Джереми Бентам ([176], 1823) и Джон Стюарт Милль ([1576], 1863) поддерживали идею ▶**утилитаризма**: рациональное принятие решений на основе максимизации полезности должно применяться во всех сферах человеческой деятельности, в том числе в области государственных политических решений, принимаемых от имени многих людей. Утилитаризм является одной из форм ▶**консеквенциализма**, основная идея которого такова: что считать правильным или неправильным, определяется ожидаемыми результатами действия.

В противоположность этому Иммануил Кант в 1875 году предложил свою теорию ▶**деонтологической этики**, базирующейся на системе установленных правил. Согласно ее положениям правильность действия определяется не по результатам, а по соответствию универсальным социальным законам, регулирующим допустимость действий, таким как “не лги” или “не убивай”. Таким образом, последователь утилитаризма имеет право на “белую” ложь, если ее ожидаемые хорошие следствия перевешивают плохие, тогда как для приверженца этики Канта это недопустимо, поскольку ложь в самой своей сути является действием неправильным. Милль признавал значение правил, но понимал их как эффективные процедуры принятия решений, составленные на результатах первичных рассуждений о последствиях. Во многих современных системах ИИ применяется именно этот подход.

1.2.2. Математика

- Каковы формальные правила формирования правильных заключений?
- Что может быть вычислено?
- Как проводить рассуждения на основе недостоверной информации?

Философы сформулировали наиболее важные идеи искусственного интеллекта, но для его преобразования в формальную науку потребовалось достичь определенного уровня математической формализации в области логики, теории вероятности и разработки новой ветви математики: теории вычислений.

Истоки идей ▶**формальной логики** можно найти уже в работах философов Древней Греции, Индии и Китая, но ее становление как математической дисциплины фактически началась с трудов Джорджа Буля (1815–1864), который детально разработал логику высказываний, или булеву логику. В 1879 году Готтлоб Фреге (1848–1925) расширил булеву логику для включения в нее объектов и отношений

[775], создав логику первого порядка, которая в настоящее время используется как наиболее фундаментальная система представления знаний.[5] Помимо своей центральной роли в ранний период исследований в области ИИ, логика первого порядка мотивировала работы Гёделя и Тьюринга, которые заложили теоретические основы вычислительной техники, как это будет объяснено ниже.

▶ **Теорию вероятности** можно рассматривать как обобщение логики на ситуации с неопределенной информацией — весьма важный вклад в теорию искусственного интеллекта. Итальянский математик Джероламо Кардано (1501–1576) первым сформулировал идею вероятности, описывая ее в терминах результатов событий с несколькими исходами, возникающих в азартных играх. В 1654 году Блез Паскаль (1623–1662), в письме Пьеру Ферма (1601–1665), показал, как можно предсказать будущее в бесконечной азартной игре и распределить средний выигрыш между игроками. Вероятность быстро стала неотъемлемой частью всех количественных наук, помогая справляться с неточностью измерений и незавершенностью теорий. Якоб Бернулли (1654–1705, дядя Даниила Бернулли), Пьер Лаплас (1749–1827), и другие внесли большой вклад в эту теорию и ввели новые статистические методы. Томас Байес (1702–1761) предложил правило обновления вероятностей с учетом новых фактов. Правило Байеса и возникшее на его основе научное направление, называемое байесовским анализом, являются важными инструментами для систем ИИ.

Формализации вероятности, в сочетании с доступностью данных, привели к появлению ▶ **статистики** как нового поля научных исследований. Одним из первых достижений в этой области стал выполненный Джоном Граунтом анализ данных переписи населения Лондона 1662 года. Первым современным статистиком считается Рональд Фишер. Он объединил идеи вероятности, планирования эксперимента, анализа данных и вычислений. В 1919 году он настаивал на том, что не смог бы выполнять свою работу без механического калькулятора под названием MILLIONAIRE (первый арифмометр, позволявший выполнять операцию умножения), даже несмотря на то, что стоимость этого калькулятора была больше, чем его годовая зарплата.

История вычислений так же стара, как история чисел, но первым нетривиальным ▶ **алгоритмом** считается алгоритм вычисления наибольшего общего знаменателя, предложенный Евклидом. Само слово “алгоритм” пришлом к нам от Мухаммеда ибн Мусы аль-Хорезми, среднеазиатского математика IX столетия, чьи труды также познакомили Европу с арабскими цифрами и алгеброй. Буль и другие ученые широко обсуждали алгоритмы логического вывода, а к концу XIX столетия даже предпринимались усилия по формализации общих принципов проведения математических рассуждений как логического вывода.

---

[5] Предложенная Готтлобом Фреге система обозначений для логики первого порядка, представлявшая собой загадочную комбинацию из текстовых и геометрических элементов, так и не нашла широкого распространения.

Курт Гёдель (1906–1978) показал, что существует эффективная процедура доказательства любого истинного высказывания в логике первого порядка Фреге и Рассела, но при этом логика первого порядка не позволяет выразить принцип математической индукции, необходимый для представления натуральных чисел. В 1931 году Гёдель показал, что действительно существуют реальные пределы вычислимости. Предложенная им ►**теорема о неполноте** показывает, что в любой теории, достаточно выразительной для описания свойств арифметики Пеано (элементарной теории натуральных чисел), существуют истинные высказывания, которые являются недоказуемыми в рамках этой теории.

Этот фундаментальный результат также может быть интерпретирован как демонстрация того, что некоторые функции на целых числах не могут быть представлены с помощью какого-либо алгоритма, т.е. они не могут быть вычислены. Это побудило Алана Тьюринга (1912–1954) попытаться точно охарактеризовать, какие функции *являются* ►**вычислимыми**, т.е. могут быть вычислены с использованием некоторой эффективной процедуры. Тезис Черча–Тьюринга предлагает отождествить общее понятие вычислимости с функциями, вычисляемыми машиной Тьюринга. Тьюринг также показал, что существуют некоторые функции, которые ни одна машина Тьюринга не может вычислить. Например, никакая машина не сможет *в общем случае* определить, будет ли указанная программа возвращать результат и прекращать работу при указанных данных или будет работать бесконечно.

Хотя понятие вычислимости очень важно для понимания возможностей вычисления, гораздо большее влияние на развитие искусственного интеллекта оказало понятие ►**разрешимости**. Грубо говоря, задача называется неразрешимой, если время, требуемое для решения отдельных примеров этой задачи, растет экспоненциально с увеличением размеров этих примеров. Различие между полиномиальным и экспоненциальным ростом сложности было впервые подчеркнуто в середине 1960-х годов в работах Кобхэма и Эдмондса. Это важно, потому что экспоненциальный рост сложности означает, что даже умеренно большие примеры могут оказаться неразрешимыми за какое-либо разумное время.

Теория ►**NP-полноты**, впервые предложенная Стивеном Куком и Ричардом Карпом, предоставляет необходимую основу для анализа разрешимости задач: любой класс задач, к которому может быть сведен класс NP-полных задач, является, по-видимому, неразрешимым. (Хотя еще не было доказано, что NP-полные задачи обязательно являются неразрешимыми, большинство теоретиков считают, что дело обстоит именно так.) Эти результаты контрастируют с тем оптимизмом, с которым в популярных периодических изданиях приветствовалось появление первых компьютеров под такими заголовками, как “Электронные супермозги”, которые думают “быстрее Эйнштейна!” Несмотря на постоянное повышение быстродействия компьютеров, экономное использование ресурсов и вынужденное несовершенство являются характерными особенностями интеллектуальных систем. Грубо говоря, наш мир — это *чрезвычайно* крупный экземпляр задачи.

1.2.3. Экономика

- Как нам следует принимать решения в соответствии с нашими предпочтениями?
- Как это следует делать, когда другие могут препятствовать нам?
- Как действовать в таких случаях, когда вознаграждение может быть получено лишь в отдаленном будущем?

Экономика как наука возникла в 1776 году, когда шотландский философ Адам Смит (1723–1790) опубликовал свою книгу *Исследование о природе и причинах богатства народов*. Смит предложил рассматривать экономику как состоящую из множества индивидуальных агентов, стремящихся к достижению собственных интересов. Смит, однако, не рассматривал стремление к финансовому обогащению как основную моральную установку: свою раннюю книгу *Теория моральных отношений* (1759) он начал с указания, что беспокойство о благополучии других является важным компонентом интересов каждого индивида.

Большинство людей считают, что экономика имеет дело исключительно с деньгами, и действительно, первый математический анализ принятия решений в условиях неопределенности — формула максимальной ожидаемой стоимости Арнольда — имел отношение к денежной стоимости ставок. Даниил Бернулли отметил, что эта формула, похоже, плохо работает в случае больших денежных сумм, например инвестиций в морские торговые экспедиции. Вместо нее он предложил принцип, построенный на максимизации ожидаемой полезности, и объяснил выбор инвестиций людьми исходя из предположения, что минимальная полезность дополнительного количества денег уменьшается, когда человек получает больше денег.

Леон Вальрас дал более общую математическую трактовку теории полезности в терминах предпочтений между азартными играми на любые результаты (не только денежные). Эта теория была улучшена Фрэнком Рамсеем, а затем усовершенствована Джоном фон Нейманом и Оскаром Моргенштерном в книге *Теория игр и экономического поведения* ([2282], 1944). Сейчас экономика уже не рассматривается как наука о деньгах — скорее, это изучение намерений и предпочтений людей.

▶ **Теория принятия решений**, объединяющая в себе теорию вероятностей и теорию полезности, предоставляет формальную и полную инфраструктуру для принятия решений (в области экономики или в другой области) в условиях неопределенности, т.е. в тех случаях, когда среда, в которой действует лицо, принимающее решение, наиболее адекватно может быть представлена лишь с помощью вероятностных описаний. Она хорошо подходит для “крупных” экономических образований, где каждый агент не обязан учитывать действия других агентов как индивидуумов. Однако в “небольших” экономических образованиях ситуация в большей степени напоминает **игру**, поскольку действия одного игрока могут существенно повлиять на полезность действий другого (или положительно, или

отрицательно). **Теория игр**, разработанная фон Нейманом и Моргенштерном, позволяет сделать неожиданный вывод: в некоторых играх рациональный агент должен действовать случайным образом или по крайней мере таким образом, который кажется случайным для соперников. В отличие от теории принятия решений, теория игр не предлагает однозначного рецепта для выбора действий. В области исследований искусственного интеллекта решения с участием нескольких агентов исследуются под заголовком **мультиагентные системы** (глава 18).

Экономисты за немногими исключениями не стремятся найти ответ на третий вопрос, приведенный в начале раздела, т.е. не предпринимают попыток выработать способ принятия рациональных решений в таких условиях, когда вознаграждение в ответ на определенные действия не предоставляется немедленно, а становится результатом нескольких действий, выполненных в определенной *последовательности*. Изучению этой темы посвящена область ▶**исследования операций**, которая возникла во время второй мировой войны в результате усилий, предпринятых в Великобритании в отношении оптимизации работы радарных установок, а в дальнейшем нашла применение и в гражданском обществе при выработке сложных управленческих решений. В работе Ричарда Беллмана ([169], 1957) формализован определенный класс последовательных задач выработки решений, называемых **марковскими процессами принятия решений**, которые рассматриваются в главе 17, а под названием **обучение с подкреплением** — в главе 22.

Работы в области экономики и исследования операций оказали большое влияние на сформулированное в этой книге понятие рациональных агентов, однако в течение многих лет исследования в области искусственного интеллекта проводились совсем по другим направлениям. Одной из причин этого была кажущаяся сложность задачи выработки рациональных решений. Тем не менее один из первых исследователей в области искусственного интеллекта, Герберт Саймон (1916–2001), получил в 1978 году Нобелевскую премию по экономике за свои ранние работы, в которых показал, что модели, основанные на ▶**разумной достаточности** (т.е. на принятии решений, которые являются "достаточно приемлемыми", вместо проведения трудоемких расчетов с целью нахождения оптимального решения), дают лучшее описание фактического поведения человека. С 1990-х годов отмечается возрождение интереса к использованию методов теории принятия решений в применении к системам искусственного интеллекта.

1.2.4. Нейронауки

- Как информация обрабатывается в мозгу?

▶**Нейронауки** — это область научных исследований, посвященная изучению нервной системы, в особенности мозга. Хотя точный способ, посредством которого мозг реализует мышление, все еще является одной из самых больших тайн в науке, тот факт, что он действительно обеспечивает мышление, был известен

в течение тысяч лет, поскольку существовали свидетельства, что сильные удары по голове могут привести к умственной недееспособности. Также давно было известно, что человеческий мозг чем-то отличается от мозга других живых существ, — примерно в 335 г. до н.э. Аристотель писал: “Из всех животных у человека самый большой мозг в отношении к его размерам”.[6] Тем не менее широкое признание того, что мозг является вместилищем сознания, пришло только в середине XVIII столетия. До этого времени в качестве возможных источников сознания рассматривались сердце и селезенка.

Поль Брока (1824–1880) в 1861 году провел исследования афазии (нарушения речи) у пациентов с повреждением мозга, которые стали отправной точкой для проведения исследований функциональной организации мозга, благодаря выявлению определенной области в левом полушарии — теперь ее называют зоной Брока, — которая отвечает за организацию речевой активности.[7] К тому времени уже было известно, что мозг состоит из нервных клеток, или ►**нейронов**, но только в 1873 году Камилло Гольджи (1843–1926) сумел разработать надежный метод, позволяющий наблюдать за отдельными нейронами в мозгу (рис. 1.1). Именно этот метод использовал Сантьяго Рамон-и-Кахаль (1852–1934) в своих пионерских исследованиях нейронных структур мозга.[8] В настоящее время общепризнано, что когнитивные функции являются результатом электрохимического воздействия этих структур. То есть ➨*совокупность простых клеток может привести к мышлению, действию и пониманию*. По содержательным словам Джона Сирла ([2025], 1992), *мозг является причиной разума*.

Теперь ученые располагают некоторыми данными о том, как связаны между собой отдельные области мозга и те части тела, которыми они управляют или от которых получают сенсорные данные. Удивительно, что подобная привязка может коренным образом измениться в течение нескольких недель, а у некоторых животных, по-видимому, имеется несколько вариантов такой привязки. Более того, еще не совсем понятно, как другие области могут взять на себя функции поврежденных областей. К тому же почти полностью отсутствуют обоснованные теории того, как осуществляется хранение информации в памяти индивидуума или как работают высокоуровневые когнитивные функции.

---

[6] С течением времени было обнаружено, что у землеройки и некоторых видов птиц соотношение веса мозга и тела превышает таковое у человека.

[7] В качестве возможного более раннего источника многие цитируют работу Александра Гуда (1824).

[8] Гольджи упорно отстаивал свое мнение, что функции мозга осуществляются в основном непрерывной средой, в которую включены нейроны, тогда как Кахаль предлагал “нейронную доктрину”. Эти ученые совместно получили Нобелевскую премию в 1906 году, но в роли лауреатов произнесли речи, содержащие взаимные антагонистичные выпады.

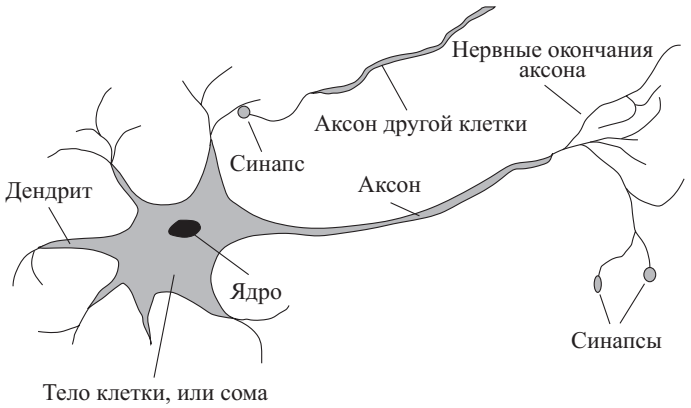

Рис. 1.1. Части нервной клетки, или нейрона. Каждый нейрон состоит из тела клетки (или сомы), которое содержит ядро клетки. От тела клетки ответвляется множество коротких волокон, называемых дендритами, и одно длинное волокно, называемое аксоном. Аксон растягивается на большое расстояние, намного превышающее то, что показано в масштабе этого рисунка. Обычно аксоны имеют длину 1 см (что в 100 раз больше диаметра тела клетки), но могут достигать 1 метра. Нейрон соединяется с другими нейронами, количество которых может составлять от 10 до 100 000 в точках соединения, называемых синапсами. Сигналы распространяются от одного нейрона к другому с помощью сложной электрохимической реакции. Эти сигналы управляют активностью мозга в течение короткого интервала времени, а также вызывают долговременные изменения состояния самих нейронов и их соединений. Считается, что именно эти механизмы служат в мозгу основой для обучения. Обработка информации происходит главным образом в коре головного мозга, которая представляет собой внешний слой нейронов мозга. По-видимому, основной ее структурной единицей является колонка ткани или *модуль*, имеющий диаметр около 0,5 мм и протяженность на всю глубину коры, толщина которой в человеческом мозгу составляет около 4 мм. Каждый модуль содержит примерно 20 000 нейронов

Первые измерения активности неповрежденного мозга начали проводить в 1929 году с изобретением Гансом Бергером электроэнцефалографа (ЭЭГ). Разработки в области получения изображений на основе функционального магнитного резонанса (МРТ) позволили нейрологам получать исключительно подробные изображения активности мозга, что обеспечило возможность проведения измерений, связанных с происходящими познавательными процессами различными интересными способами. Эти возможности постоянно расширяются благодаря прогрессу в области регистрации нейронной активности отдельной клетки и в методах ►**оптогенетики**, позволяющих как измерять, так и управлять активностью отдельных нейронов, модифицированных таким образом, чтобы сделать их чувствительными к свету.

Разработка ▶**интерфейсов "мозг–машина"** (Лебедев и Николелис [1369], 2006) для сенсорного и двигательного управления не только обещает получить средства восстановления функциональных возможностей людей с ограниченными возможностями, но также проливает свет на многие аспекты нейронных систем. Замечательный вывод этой работы состоит в том, что мозг способен настроить себя на успешное взаимодействие с внешним устройством, рассматривая его в действии как еще один орган чувств или конечность.

Мозг и цифровые компьютеры имеют несколько различающиеся свойства. На рис. 1.2 показано, что продолжительность рабочего цикла современного компьютера в миллион раз меньше, чем мозга. Мозг компенсирует эту разницу за счет гораздо большего объема памяти и количества взаимосвязей в сравнении с персональным компьютером даже самого высокого класса, хотя крупнейшие суперкомпьютеры по некоторым показателям уже соответствуют нашему мозгу. Футуристы придают большое значение этим показателям, указывая на приближение ▶**особого момента**, когда компьютеры достигнут сверхчеловеческого уровня производительности (Винге [2272], 1993; Курцвейл [1330], 2005; Докторов и Стросс [629], 2012), а затем стремительно улучшат себя в еще большей степени. Однако сравнение самих числовых значений недостаточно информативно. Даже в случае создания компьютера с практически неограниченной мощностью, нам по-прежнему потребуются дальнейшие концептуальные прорывы в нашем понимании интеллекта (см. главу 28). Грубо говоря, без наличия правильной теории более быстрые машины будут просто быстрее давать нам неправильные ответы.

	Суперкомпьютер	Персональный компьютер	Мозг человека
Вычислительные модули	10^6 ГП + ЦП	8-ядерный ЦП	10^6 колонок
	10^{15} транзисторов	10^{10} транзисторов	10^{11} нейронов
Модули памяти	10^{16} байт ОП	10^{10} байт ОП	10^{11} нейронов
	10^{17} байт ДП	10^{12} байт ДП	10^{14} синапсов
Цикл обработки	10^{-9} секунды	10^{-9} секунды	10^{-3} секунды
Операций в секунду	10^{18}	10^{10}	10^{17}

Рис. 1.2. Грубое сравнение характеристик лидирующего суперкомпьютера Summit (Фельдман [719], 2017), типичного персонального компьютера образца 2019 года, и человеческого мозга. (Здесь ГП — графический процессор, ЦП — центральный процессор, ОП — оперативная память, ДП — дисковая память.) Мощность человеческого мозга не изменялась сколько-нибудь значительно на протяжении тысяч лет, тогда как мощность суперкомпьютеров возросла от мегафлопсов в 1960-х до гигафлопсов в 1980-х, терафлопсов в 1990-х, петафлопсов в 2008 году и эксафлопсов в 2018 году (1 эксафлопс = 10^{18} операций с плавающей точкой в секунду)

1.2.5. Психология

- Как люди и животные думают и действуют?

Истоки научной психологии обычно прослеживаются до работ немецкого физика Германа фон Гельмгольца (1821–1894) и его студента Вильгельма Вундта (1832–1920). Гельмгольц применил научный метод для изучения зрения человека. Выпущенная им книга *Справочник по физиологической оптике* даже в наши дни характеризуется как "непревзойденный по своей важности вклад в изучение физики и физиологии зрения человека" (Нальва [1655], 1993). В 1879 году Вундт открыл первую лабораторию по экспериментальной психологии в Лейпцигском университете. Он настаивал на проведении тщательно контролируемых экспериментов, в которых его сотрудники выполняли задачи по восприятию или формированию ассоциаций, проводя интроспективные наблюдения за своими мыслительными процессами. Тщательный контроль позволил ему сделать очень многое для превращения психологии в науку, но субъективный характер делал маловероятным, что экспериментаторы когда-либо будут опровергать их собственными теориями.

С другой стороны, биологам, изучающим поведение животных, интроспективные данные недоступны, поэтому они разработали объективную методологию, которую Г.С. Дженнингс описал в своей нашумевшей работе *Поведение низших организмов* ([1133], 1906). Распространив этот подход на людей, сторонники движения, возглавляемого Джоном Уотсоном (1878–1958) и получившего название ▶**бихевиоризм**, отвергали любую теорию, учитывающую мыслительные процессы, на том основании, что интроспекция не может предоставлять надежные свидетельства. Бихевиористы настаивали на изучении только объективных мер восприятия (или *стимулов*), предъявленных животному, и их вытекающих из этого действий (или *откликов*). Бихевиоризм позволил многое узнать о крысах и голубях, но оказался менее успешным при изучении поведения человека.

Взгляды, согласно которым мозг рассматривается как устройство обработки информации, характерны для представителей ▶**когнитивной психологии** и прослеживаются по крайней мере до работ Уильяма Джеймса (1842–1910). Гельмгольц также настаивал на том, что восприятие связано с определенной формой подсознательного логического вывода. В Соединенных Штатах такой подход к изучению познавательных процессов был в основном отвергнут из-за широкого распространения бихевиористских взглядов, но на факультете прикладной психологии Кембриджского университета, возглавляемом Фредериком Бартлеттом (1886–1969), когнитивное моделирование активно поддерживалось. В книге *Природа объяснения* ([491], 1943) ученик и последователь Бартлетта Кеннет Крэйк привел весомые доводы в пользу допустимости применения таких "мыслительных" терминов, как "убеждения" и "цели", доказав, что они являются не менее научными, чем, скажем, такие термины, применяемые в рассуждениях о газах, как "давление" и "температура", несмотря на то что речь в них идет о молекулах, которые сами не обладают этими характеристиками.

Крэйк обозначил следующие три этапа деятельности агента, основанного на знаниях: во-первых, действующий стимул должен быть преобразован во внутреннее представление, во-вторых, с этим представлением должны быть выполнены манипуляции с помощью познавательных процессов для выработки новых внутренних представлений и, в-третьих, они должны быть, в свою очередь, снова преобразованы в действия. Он наглядно объяснил, почему такой проект является приемлемым для любого агента.

> Если живой организм несет в своей голове "модель в уменьшенном масштабе" внешней реальности и своих возможных действий, то он способен проверять различные альтернативы, делать заключение, какая из них является наилучшей для него, реагировать на будущие ситуации, прежде чем они возникнут, использовать знания о прошлых событиях, сталкиваясь с настоящим и будущим, и во всех отношениях реагировать на опасности, встречаясь с ними, гораздо полнее, безопаснее и более компетентно (Крейк [491], 1943).

В 1945 году, после смерти Крэйка в результате аварии во время поездки на велосипеде, его работа была продолжена Дональдом Броудбентом, чья книга *Восприятие и коммуникация* ([310], 1958) была одной из первых работ по моделированию психологических явлений как процессов обработки информации. Между тем в Соединенных Штатах работы в области компьютерного моделирования привели к созданию такого научного направления, как ▶**когнитивная наука** (или **когнитивистика**). Зарождение этого направления исследований произошло на одном из семинаров в Массачусетсском технологическом институте в сентябре 1956 года. (Ниже будет показано, что это событие произошло всего лишь через два месяца после проведения конференции, на которой "родился" сам искусственный интеллект.)

На этом семинаре Джордж Миллер представил доклад *Волшебное число "семь"*, Ноам Хомский прочитал доклад *Три модели языка*, а Аллен Ньюэлл и Герберт Саймон представили свою работу *Машина теории логики*. В этих трех работах, получивших широкое признание, было показано, как можно использовать компьютерные модели для решения задач в области психологии памяти, обработки естественного языка и логического мышления соответственно. В настоящее время среди психологов находит широкое признание (хотя и не является универсальным) взгляд, что "любая теория познания должна быть подобна компьютерной программе" (Андерсон [47], 1980), т.е. должна описывать работу любой познавательной функции в терминах обработки информации.

Исходя из целей данного обзора, мы не будем рассматривать область **взаимодействия человека с компьютером** как раздел психологии. Дуг Энгельбарт, один из пионеров в этой области, отстаивал идею ▶**усиления интеллекта** (УИ) вместо ИИ. Он считал, что компьютеры должны *расширять* человеческие способности, а не автоматизировать решение задач человека. В 1968 году Энгельбарт в своем выступлении на компьютерной конференции (позднее его окрестили как "Мать всех демонстраций") впервые продемонстрировал такие невероятные на то время

новшества, как компьютерная мышь, оконный компьютерный интерфейс, гипертекстовый документ и даже проведение видеоконференции с коллективной работой над одним документом. Все это имело единственную цель — продемонстрировать, чего способны коллективно достичь занятые умственным трудом люди с помощью различных средств, отвечающих концепции усиления интеллекта.

Сегодня мы склонны рассматривать концепции УИ и ИИ как две стороны одной медали, делая акцент в первом случае на контроле со стороны человека, а во втором — на интеллектуальном поведении со стороны машины. И то, и другое необходимо, чтобы машины были полезны людям.

1.2.6. Компьютерная техника

- Как можно создать эффективный компьютер?

Современный цифровой электронный компьютер был изобретен независимо и почти одновременно учеными трех стран, сражавшихся во второй мировой войне. Первым *действующим* компьютером был электромеханический компьютер "Хит Робинсон",[9] построенный в 1943 году командой Алана Тьюринга с единственной целью: расшифровка перехваченных сообщений немецких войск. Позднее в том же 1943 году эта же группа разработала Colossus — мощную машину универсального назначения на основе электронных ламп.[10] Первым работающим *программируемым* компьютером был Z-3, изобретенный Конрадом Цузе в Германии в 1941 году. Цузе также изобрел способ представления чисел с плавающей точкой и первый язык программирования высокого уровня, *Plankalkül*. Первый *электронный* компьютер, ABC, был собран Джоном Атанасовым и его студентом Клиффордом Берри между 1940 и 1942 годами в Университете штата Айова. Работа Атанасова не получила необходимой поддержки или признания в отличие от проекта ENIAC — компьютера, созданного в рамках засекреченного военного проекта в Университете штата Пенсильвания командой разработчиков, включавшей Джона Мокли и Дж. Преспера Эккерта. Именно этот проект оказался самым влиятельным предтечей современных компьютеров.

С тех пор появление нового поколения компьютерной техники всегда сопровождалось увеличением скорости процессоров и емкости памяти с одновременным снижением стоимости — тенденция, зафиксированная в ▶**законе Мура**. Согласно этому закону производительность компьютеров удваивается каждые 18 месяцев или около того. Так продолжалось до 2005 года, когда проблема отвода тепла вынудила производителей обратиться к увеличению количества ядер в процессорах

[9] Сложная машина, названная в честь британского карикатуриста, изображавшего причудливые и нелепо сложные способы решения повседневных задач, например приготовления тостов.

[10] В послевоенный период Тьюринг хотел использовать эти компьютеры для исследований в области ИИ, например он создал схему первой шахматной программы (Тьюринг и др. [2236] 1953), но британское правительство запретило проведение этих исследований.

вместо повышения их тактовой частоты. В настоящее время ожидается, что дальнейшее увеличение функциональности будет достигаться за счет интенсивного распараллеливания вычислительных процессов — здесь просматривается любопытная схожесть со свойствами мозга. Также можно отметить новые аппаратные разработки, основанные на идее, что, имея дело с неопределенным миром, нам не требуется числовая точность, обеспечиваемая форматом слова в 64 бита. Вполне достаточно будет всего 16 бит (как в формате `bfloat16`) или даже 8 бит, что существенно ускорит обработку.

Мы лишь совсем недавно смогли увидеть компьютерное оборудование, специально предназначенное для применения в области искусственного интеллекта. Это, например, графические процессоры (ГПУ), тензорные процессоры (ТПУ) и суперпроцессор параллельной обработки Wafer Scale Engine (WSE). В период с 1960-х до 2012 года объем вычислительной мощности, используемой для обучения высокопроизводительных машин, соответствовал закону Мура. Начиная с 2012 года ситуация изменилась: с 2012 по 2018 год суммарные показатели выросли в 300 000 раз (!), что соответствует их удваиванию каждые 100 дней или около того (Амодей и Эрнандес [44], 2018). Модель машинного обучения, которая в 2014 году требовала на выполнение полного рабочего дня, в 2018 году выполнялась всего за две минуты (Йинг и др. [2404], 2018). Хотя ▶**квантовые вычисления** пока еще не достигли уровня практического применения, этот подход обещает гораздо большее ускорение для некоторых важных подклассов алгоритмов ИИ.

Безусловно, вычислительные устройства существовали и до появления электронного компьютера. Первые автоматизированные устройства, появившиеся еще в XVII столетии, уже упоминались в разделе 1.2.1. Первым *программируемым* устройством был ткацкий станок, изобретенный в 1805 году Жозефом Марией Жаккардом (1752–1834), в котором для хранения инструкций по плетению узоров ткани использовались перфокарты.

В середине XIX века Чарльз Бэббидж (1792–1871) разработал две вычислительные машины, ни одну из которых он так и не завершил. Его дифференциальная машина была предназначена для вычисления математических таблиц для инженерных и научных расчетов. Значительно позднее, уже в 1991 году, она все же была построена, а ее работоспособность наглядно продемонстрирована (Свейд [2160], 2000). Вторая, аналитическая, машина Бэббиджа была гораздо более амбициозной: она включала адресацию памяти, сохраненные программы на перфорированных картах Жаккарда, команды условных переходов. Это была первая машина, способная выполнять универсальные вычисления.

Коллега Бэббиджа Ада Лавлейс, дочь поэта лорда Байрона, осознала потенциал этого устройства, описав его как “мыслящая или… рассуждающая машина”, способная рассуждать обо “всех предметах во Вселенной” (Лавлейс [1447], 1843). Она также предвидела возможность шумихи вокруг ИИ, написав, что “желательно защититься от возможности появления преувеличенных идей, которые могут возникнуть в отношении мощности аналитической машины”. К сожалению, машины Бэббиджа и идеи Ады Лавлейс были в значительной степени забыты.

Искусственный интеллект во многом обязан также тем направлениям компьютерных наук, которые связаны с созданием операционных систем, языков программирования и прочих инструментальные средств, необходимых для написания современных программ (и статей о них). Но эта область научной деятельности является также одной из тех, в которых искусственный интеллект в полной мере возмещает свои долги: работы в области искусственного интеллекта стали источником многих идей, которые затем были воплощены в основных направлениях развития компьютерных наук, включая разделение времени, интерактивные интерпретаторы, персональные компьютеры с оконными интерфейсами и поддержкой позиционирующих устройств, применение эффективных сред разработки, создание типов данных в виде связных списков, автоматическое управление памятью и ключевые концепции символического, функционального, декларативного и объектно-ориентированного программирования.

1.2.7. Теория управления и кибернетика

- Как артефакты могут работать под собственным управлением?

Ктесибий из Александрии (ок. 250 до н.э.) построил первую самоуправляемую машину: водяные часы с регулятором, поддерживающим постоянную скорость потока. Это изобретение изменило определение того, что может делать артефакт. Ранее только живые существа могли изменять свое поведение в ответ на изменения в окружающей среде. Другие примеры саморегулирующихся систем управления с обратной связью включают регулятор парового двигателя, созданный Джеймсом Уаттом (1736–1819), и термостат, изобретенный Корнелисом Дреббелем (1572–1633), который также изобрел подводную лодку. Джеймс Клерк Максвелл ([1518], 1868) положил начало математической теории систем управления.

В послевоенный период центральной фигурой в разработке ▶**теории управления**, был Норберт Винер (1894–1964). Винер был блестящим математиком, который работал со многими учеными, включая Бертрана Рассела, прежде чем у него появился интерес к изучению биологических и механических систем управления и их связи с познанием. Как и Крэйк (который также использовал системы управления в качестве психологических моделей), Винер и его коллеги Артуро Розенблют и Джулиан Бигелоу бросили вызов ортодоксальным бихевиористским взглядам ([1914], 1943). Они рассматривали целенаправленное поведение как обусловленное действием некого регуляторного механизма, пытающего минимизировать "ошибку" — различие между текущим и целевым состояниями. В конце 1940-х годов Винер совместно с Уорреном Мак-Каллоком, Уолтером Питтсом и Джоном фон Нейманом организовал ряд влиятельных конференций, на которых рассматривались новые математические и вычислительные модели познания. Книга Винера ▶***Кибернетика*** ([2339], 1948) стала бестселлером и убедила широкие круги общественности в том, что мечта о создании машин, обладающих искусственным интеллектом, воплотилась в реальность.

Между тем в Великобритании У. Росс Эшби впервые *применил* подобные идеи (Эшби [82], 1940). Эшби, Алан Тьюринг, Грей Уолтер и другие сформировали "Ratio Club" для "тех, кто разделял идеи Винера до появления книги Винера". В книге *Дизайн для мозга* ([83], 1948; [84], 1952) Эшби разрабатывал свою идею о том, что разум может быть создан при использовании ▶**гомеостатических** устройств, содержащих соответствующие петли обратной связи, обеспечивающие достижение стабильного адаптивного поведения.

Современная теория управления, особенно ее ветвь с названием *стохастическое оптимальное управление*, ставит своей целью проектирование систем, которые максимизируют ▶**целевую функцию** во времени. Это примерно соответствует представлению авторов настоящей книги об искусственном интеллекте как о проектировании систем, которые действуют оптимальным образом. Почему же в таком случае искусственный интеллект и теория управления рассматриваются как две разные научные области, особенно если учесть, какие тесные взаимоотношения связывали их основателей? Ответ на этот вопрос состоит в том, что существует также тесная связь между математическими методами, которые были знакомы участникам этих разработок, и соответствующими множествами задач, которые были охвачены в каждом из этих подходов к описанию мира. Дифференциальное и интегральное исчисление, а также матричная алгебра, являющиеся инструментами теории управления, в наибольшей степени подходят для анализа систем, которые могут быть описаны с помощью фиксированных множеств непрерывно изменяющихся переменных, тогда как сама область исследований ИИ была отчасти основана как способ избежать этих ограничений математических средств. Такие инструменты, как логический вывод и вычисления, позволили исследователям искусственного интеллекта успешно рассматривать некоторые проблемы, такие как понимание естественного языка, зрение и символическое планирование, полностью выходящие за рамки исследований, предпринимавшихся теоретиками в области теории управления.

1.2.8. Лингвистика

- Каким образом язык связан с мышлением?

В 1957 году Б.Ф. Скиннер опубликовал свою книгу *Вербальное поведение*. Это был всеобъемлющий, подробный отчет о результатах исследований по изучению языка, проведенных в рамках бихевиористского подхода, написанный наиболее выдающимся экспертом в этой области. Но весьма любопытно то, что рецензия к этой книге стала не менее известной, чем сама книга, и послужила причиной почти полного исчезновения интереса к бихевиоризму. Автором этой рецензии был Ноам Хомский, который сам только что опубликовал книгу *Синтаксические структуры* ([422], 1957) с изложением собственной теории. Хомский показал, что бихевиористская теория не позволяет понять истоки творческой деятельности, осуществляемой с помощью языка, — она не объясняет, почему ребенок способен

понимать и складывать предложения, которые он до сих пор никогда еще не слышал. Теория Хомского, основанная на синтаксических моделях, восходящих к работам древнеиндийского лингвиста Панини (примерно 350 год до н.э.), позволяла объяснить этот феномен и, в отличие от предыдущих теорий, оказалась достаточно формальной для того, чтобы ее можно было реализовать в виде программ.

Таким образом, современная лингвистика и искусственный интеллект, которые "родились" примерно в одно и то же время и продолжают расти вместе, пересекаются в гибридной области, называемой ▶**вычислительная лингвистика** или **обработка естественного языка**. Со временем было обнаружено, что проблема понимания языка является гораздо более сложной, чем это казалось в 1957 году. Для понимания языка требуется понимание предмета и контекста речи, а не только анализ структуры предложений. Это утверждение теперь кажется очевидным, но сам данный факт не был широко признан до 1960-х годов. Основная часть ранних работ в области **представления знаний** (науки о том, как преобразовать знания в такую форму, которой может оперировать компьютер) была привязана к языку и подпитывалась исследованиями в области лингвистики, которые, в свою очередь, основывались на результатах философского анализа языка, проводившегося в течение многих десятков лет.

1.3. История искусственного интеллекта

Один из быстрых способов подвести итог вехам в истории ИИ состоит в том, чтобы перечислить некоторых лауреатов премии Тьюринга: Марвин Мински (1969) и Джон Мак-Карти (1971) — за формирование основ этого научного направления, построенных на представлении и рассуждении; Эдвард Фейгенбаум и Радж Редди (1994) — за разработку экспертных систем, кодирующих знания человека для решения реальных проблем; Джуди Перл (2011) — за разработку вероятностных методов рассуждения, которые принципиальным образом решают проблему неопределенности; и наконец, Йэшуа Бегио, Джеффри Хинтон и Янн ЛеКун (2019) — за разработку концепции "глубокого обучения" (многослойные нейронные сети), критически важной составляющей современной вычислительной технологии. В остальной части этого раздела каждая фаза истории ИИ описывается более подробно.

1.3.1. Начальный этап развития искусственного интеллекта (1943–1956)

Первая работа, которую сейчас по общему признанию относят к искусственному интеллекту, была выполнена Уорреном Мак-Каллоком и Уолтером Питтсом ([1537], 1943). Вдохновленные работой по математическому моделированию советника Питтса Николаса Рашевского, они опирались на три источника: знание основ физиологии и функции нейронов мозга; формальный анализ логики высказываний,

взятый из работ Рассела и Уайтхеда, и теорию вычислений Тьюринга. Мак-Каллок и Питтс предложили модель из искусственных нейронов, в которой каждый нейрон мог находиться в состоянии "включено" или "выключено" и переход в состояние "включено" происходил в ответ на стимуляцию со стороны достаточного количества соседних нейронов. Состояние нейрона рассматривалось как "фактически эквивалентное высказыванию, в котором предлагается адекватное количество стимулов". В своей работе они показали, что любая вычислимая функция может быть вычислена с помощью некоторой сети из соединенных нейронов и что все логические связки ("И", "ИЛИ", "НЕ" и т.д.) могут быть реализованы с помощью простых сетевых структур. Кроме того, Мак-Каллок и Питтс выдвинули предположение, что структурированные соответствующим образом сети способны к обучению. Дональд Хебб ([997], 1949) продемонстрировал простое правило обновления для модификации количества соединений между нейронами. Предложенное им правило, называемое теперь правилом ▶ **хеббовского обучения**, продолжает служить основой для моделей, широко используемых и в наши дни.

Два студента Гарварда, Марвин Мински и Дин Эдмондс, в 1950 году создали первый компьютер на основе нейронной сети. В этом компьютере, получившем название SNARC, для моделирования сети из 40 нейронов использовалось 3000 электронных ламп плюс механизм автопилота с бомбардировщика B-24. Позже, в Принстоне, Мински изучал возможности выполнения универсальных вычислений в нейронных сетях. Когда он защищал диссертацию доктора философии, аттестационная комиссия выразила сомнение, может ли работа такого рода рассматриваться как математическая, на что фон Нейман, по словам современников, возразил: "Сегодня — нет, но когда-то будет".

Существует большое количество примеров других ранних работ, которые можно охарактеризовать как относящиеся к искусственному интеллекту, в том числе две программы игры в шашки, разработанные независимо одна от другой в 1952 году Кристофером Стрейчем в Университете Манчестера и Артуром Сэмюэлом в IBM. Однако решающим оказалось мнение Алана Тьюринга. Начиная с 1947 года он читал лекции по этому вопросу в Лондонском математическом обществе и сформулировал убедительную программу исследований в статье *Вычислительные машины и интеллект* [2235], опубликованной в 1950 году. В этой статье он описал тест Тьюринга, принципы машинного обучения, генетические алгоритмы и обучение с подкреплением. Он рассмотрел многие возражения в отношении возможностей ИИ, как это описано в главе 27. Он также предположил, что создать ИИ на уровне человеческого интеллекта будет проще путем разработки алгоритмов обучения, с последующим обучением машины, а не прямым программированием ее интеллекта вручную. В последующих лекциях он предупреждал, что достижение этой цели может быть не самым лучшим решением для человечества.

В 1955 году Джон Маккарти уговорил Марвина Мински, Клода Шеннона и Натаниэля Рочестера помочь ему собрать всех американских исследователей, проявляющих интерес к теории автоматов, нейронным сетям и исследованиям

интеллекта. В конечном итоге они организовали двухмесячный семинар в Дартмуте летом 1956 года. Всего на нем присутствовало 10 участников, включая Аллена Ньюэлла и Герберта Саймона из Института Карнеги,[11] Тренчарда Мура из Принстона, Артура Сэмюэла из компании IBM, а также Рея Соломонова и Оливера Селфриджа из Массачусетсского технологического института. В приглашении говорилось следующее.[12]

> Мы предлагаем провести двухмесячное исследование искусственного интеллекта для 10 человек летом 1956 года в Дартмутском колледже в Ганновере, штат Нью-Гемпшир. Исследование будет проводиться на основании предположения, что каждый аспект обучения или любой другой функции интеллекта в принципе может быть описан настолько точно, что это позволит создать машину, способную его имитировать. Будут предприняты попытки выяснить, как разрабатывать машины, способные использовать язык, формулировать абстракции и концепции, решать различные задачи, в настоящее время доступные только людям, а также улучшать самих себя. Мы считаем, что можно добиться значительного прогресса в одном или нескольких из указанных направлений, если тщательно отобранная группа ученых будет совместно работать над этим в течение лета.

Несмотря на этот оптимистический прогноз, дартмутский семинар не привел к появлению каких-либо новых крупных открытий. Ньюэл и Саймон представили, вероятно, наиболее зрелую работу, систему доказательства математических теорем под названием Logic Theorist (LT). Саймон заявил: “Мы изобрели компьютерную программу, способную мыслить не численно, и тем самым решили освященную веками проблему разума и тела”.[13] Вскоре после семинара эта программа смогла доказать большинство теорем из второй главы книги *Принципы математики* Рассела и Уайтхеда ([2331], 1910). Как говорили, Рассел был в восторге, когда ему сказали, что для одной из теорем LT смогла найти доказательство, которое оказалось короче, чем приведенное в его книге. Однако редактора журнала *Journal of Symbolic Logic* оказались менее впечатленными и отклонили статью, авторами которой были указаны Ньюэлл, Саймон и программа Logic Theorist.

---

[11] Теперь это учебное заведение называется “Университет Карнеги–Меллона” (Carnegie–Mellon University — CMU).

[12] Это было первое официальное использование принадлежащего Маккарти термина “искусственный интеллект”. Может быть, вариант “вычислительная рациональность” был бы более точным и менее пугающим, но “ИИ” с тех пор прочно вошел в обиход. На 50-летии Дартмутской конференции Маккарти заявил, что он избегал терминов “компьютер” и “вычислительный” в знак уважения к Норберту Винеру, который в то время пропагандировал аналоговые кибернетические устройства, а не цифровые компьютеры.

[13] В процессе создания программы LT Ньюэлл и Саймон разработали также язык обработки списков IPL. У них не было компилятора, поэтому ученые переводили программы со своего языка в машинный код вручную. Чтобы избежать ошибок, они работали параллельно, называя друг другу двоичные числа после записи каждой команды, чтобы убедиться в том, что они совпадают.

1.3.2. Ранний энтузиазм, большие ожидания (1952–1969)

Интеллектуальный истеблишмент в 1950-х годах в своем большинстве продолжал считать, что “ни одна машина не сможет выполнить действие X”. (Длинный список таких X, собранный Тьюрингом, приведен в главе 27.) Исследователи в области искусственного интеллекта, естественно, отвечали на это, демонстрируя способность машин решать одну задачу X за другой. Чаще всего они обращались к задачам, которые принято считать свидетельством интеллектуальности человека, — играм, головоломкам, математике, тестам IQ. Позднее Джон Маккарти назвал этот период эпохой “Смотри, мама, что я умею!”

За первой успешной разработкой Ньюэлла и Саймона, программой LT, последовало создание программы общего решателя задач (General Problem Solver — GPS). В отличие от программы Logic Theorist, эта программа с самого начала была предназначена для моделирования процедуры решения задач человеком. Как оказалось, в пределах того ограниченного класса головоломок, которые была способна решать эта программа, порядок, в котором она рассматривала подцели и возможные действия, был аналогичен тому подходу, который применяется людьми для решения таких же проблем. Поэтому программа GPS была, по-видимому, первой программой, в которой был воплощен подход к “организации мышления по такому же принципу, как и у человека”. Результаты успешного применения GPS и последующих программ в качестве модели познания позволили сформулировать знаменитую гипотезу ▶**физической символической системы**, в которой утверждается, что существует “физическая символическая система, которая имеет необходимые и достаточные средства для интеллектуальных действий общего вида”. Под этим подразумевается, что любая система, проявляющая интеллект (человек или машина), должна действовать по принципу манипулирования структурами данных, состоящими из символов. Ниже будет показано, что эта гипотеза во многих отношениях оказалась уязвимой для критики.

В компании IBM Натаниэль Рочестер и его коллеги создают некоторые из первых ИИ-программ. Герберт Гелернтер ([828], 1959) создал программу Geometry Theorem Prover, которая была в состоянии доказать такие геометрические теоремы, которые многие студенты-математики сочли бы достаточно сложными. Эта работа стала предшественницей современных систем доказательства математических теорем.

Из всех исследовательских работ, проделанных за этот период, возможно, самой влиятельной в долгосрочной перспективе стала разработка Артура Сэмюэла по игре в шашки. Благодаря использованию методов, которые мы сейчас называем обучением с подкреплением (см. главу 22), программы Сэмюэла в конечном итоге научились играть на уровне хорошо подготовленного любителя. Тем самым Сэмюэл опроверг утверждение, что компьютеры способны выполнять только то, чему их учили: одна из его программ быстро научилась играть лучше, чем ее создатель. Работа этой программы была показана на телевидении в 1956 году и произвела

очень сильное впечатление на зрителей. Как и Тьюринг, Сэмюэл с трудом находил машинное время. Работая по ночам, он использовал компьютеры, которые все еще находились на испытательной площадке производственного предприятия компании IBM. Программа Сэмюэла стала предшественницей более поздних систем, таких как TD-Gammon (Тезауро [2192], 1992), прочно утвердившейся среди лучших игроков в нарды в мире, и AlphaGo (Сильвер и др. [2063], 2016), потрясшей мир тем, что победила человека, чемпиона мира по игре в го (см. главу 5).

В 1958 году Джон Маккарти внес два важных вклада в развитие искусственного интеллекта. В документе MIT AI Lab Memo No. 1 он привел определение нового языка высокого уровня ▶ **Lisp**, которому суждено было стать доминирующим языком программирования для искусственного интеллекта в ближайшие 30 лет. В статье под названием *Программы со здравым смыслом* ([1529], 1958) он выдвинул концептуальное предложение о разработке систем искусственного интеллекта, основанных на знаниях и рассуждении. В этой статье он описал гипотетическую программу Advice Taker, способную воплощать общие знания о мире, а затем использовать их для выработки плана действий. Концепция была проиллюстрирована с помощью простых логических аксиом, которых было достаточно для разработки плана проезда в аэропорт. Также программа была задумана таким образом, чтобы принимать новые аксиомы в процессе работы, что позволяло ей достигать нужного уровня компетентности в других областях *без перепрограммирования*. Следовательно, в программе Advice Taker были воплощены центральные принципы представления знаний и проведения рассуждений: необходимо иметь формальное, детализированное представление о мире и его функционировании и уметь манипулировать этим представлением с помощью дедуктивных процедур. Данная статья оказала большое влияние на всю область ИИ, оставаясь вполне актуальной и сегодня.

Все тот же 1958 год также стал годом, когда Марвин Мински перешел в Массачусетсский технологический институт (МТИ). Но успешное его сотрудничество с Маккарти продолжалось недолго. Маккарти делал акцент на способах представления и проведении рассуждений в формальной логике, тогда как Мински в бо́льшей степени интересовался тем, как заставить программы работать, и в конечном итоге у него сформировалось отрицательное отношение к логике. В 1963 году Маккарти основал лабораторию искусственного интеллекта в Стандфордском университете. Разработанный им план использования логики для создания окончательной версии программы Advice Taker выполнялся быстрее, чем было задумано, благодаря открытию Дж. А. Робинсоном метода резолюции (полного алгоритма доказательства теорем для логики первого порядка; см. главу 9). Работы, выполненные в Стандфордском университете, подчеркнули важность применения методов общего назначения для проведения логических рассуждений. В число логических приложений вошли системы формирования ответов на вопросы и планирования Корделла Грина ([920], 1969), а также робототехнический проект Shakey, разрабатывавшийся в Стандфордском научно-исследовательском институте (Stanford Research

Institute — SRI). Последний проект, который подробно рассматривается в главе 26, впервые продемонстрировал полную интеграцию логических рассуждений и физической активности.

В МТИ Мински руководил работой ряда студентов, выбравших для себя небольшие задачи, для решения которых, как тогда казалось, требовалась интеллектуальность. Эти ограниченные проблемные области получили название ▶**микромиры**. Программа SAINT Джеймса Слэгла ([2083], 1963) была способна решать задачи интегрирования в закрытом исчислении, типичные для первых курсов колледжей. Программа ANALOGY Тома Эванса ([707], 1968) решала задачи выявления геометрических аналогий, используемые в тестах IQ. Программа STUDENT Дэниэла Боброва ([237], 1967) решала изложенные в виде рассказа алгебраические задачи, подобные приведенной ниже.

> Если количество заказов, полученных Томом, вдвое превышает квадратный корень из 20% опубликованных им рекламных объявлений, а количество этих рекламных объявлений равно 45, то каково количество заказов, полученных Томом?

Самым известным примером микромира был ▶**мир блоков**, состоящий из множества цельных блоков, размещенных на поверхности стола (или, чаще, имитации стола), как показано на рис. 1.3. Типичной задачей в этом мире является изменение расположения блоков определенным образом с использованием манипулятора-робота, который в каждый момент может захватывать только один блок. Мир блоков стал основой для проекта системы технического зрения Дэвида Хаффмена ([1091], 1971), работы по изучению зрения и удовлетворения ограничений Дэвида Уолтса ([2291], 1975), теории обучения Патрика Уинстона ([2362], 1970), программы понимания естественного языка Тэрри Винограда ([2361], 1972) и планировщика в мире блоков (Скотта Фалмана [709], 1974).

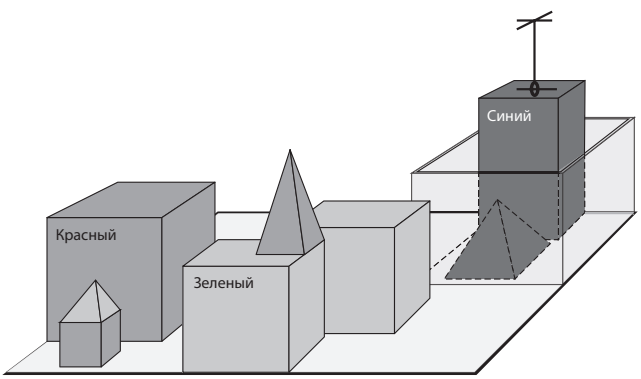

Рис. 1.3. Сцена из мира блоков. Программа SHRDLU Тэрри Винограда [2361] только что завершила выполнение команды “Найти блок, более высокий по сравнению с тем, который находится в манипуляторе, и поместить его в ящик”

Активно продвигались также исследования, основанные на ранних работах по созданию нейронных сетей Мак-Каллока и Питтса. В работе Шмуэля Винограда и Коуэна ([2360], 1963) было показано, как большое количество элементов в своей совокупности может представить отдельную концепцию с соответствующим увеличением надежности и степени распараллеливания. Методы обучения Хебба были усовершенствованы в работах Берни Видроу ([2335], 1960), ([2334], 1962), который называл свои сети **адалинами**, а также Френка Розенблатта ([1912], 1962), создателя **перцептронов**. Розенблатт доказал **теорему сходимости перцептрона**, которая подтверждает, что предложенный им алгоритм обучения позволяет корректировать количество соединений перцептрона в соответствии с любыми входными данными при условии, что такое соответствие существует.

1.3.3. Столкновение с реальностью (1966–1973)

С самого начала исследователи ИИ не стеснялись делать прогнозы в отношении своих будущих успехов. Так, в свое время часто цитировалось приведенное ниже предсказание Герберта Саймона, опубликованное им в 1957 году.

> Я не ставлю перед собой задачу удивить или шокировать вас, но проще всего я могу подвести итог, сказав, что теперь мы живем в мире машин, которые думают, учатся и создают. Более того, их способность выполнять эти действия будет продолжать быстро расти до тех пор, пока — в обозримом будущем — круг проблем, с которыми они смогут справиться, можно будет сопоставить с кругом проблем, в которых до сих пор был нужен человеческий интеллект.

Такое выражение, как “обозримое будущее”, довольно расплывчато, но Саймон сделал также более конкретные прогнозы, что через десять лет компьютер станет чемпионом мира по шахматам и что машиной будет доказана важная математическая теорема. Эти предсказания сбылись (или почти сбылись), но не через десять, а через сорок лет. Излишняя самоуверенность Саймона была обусловлена тем, что первые системы искусственного интеллекта демонстрировали многообещающую производительность, хотя и на простых примерах. Однако эти ранние системы почти всегда терпели поражение, сталкиваясь с более трудными проблемами.

У этих неудач были две основные причины. Первая заключалась в том, что многие ранние системы ИИ были основаны главным образом на “информированном самоанализе” относительно того, как люди решают задачу, а не на тщательном анализе самой задачи и выяснении, что означает решить ее и что должен делать алгоритм, чтобы с большой вероятностью привести к такому решению.

Вторая причина неудач — отсутствие правильной оценки сложности для машинной обработки многих задач, которые пытались решить с помощью ИИ. Большинство ранних систем решения проблем работали по методу проверки различных комбинаций возможных шагов, пока решение не будет найдено. Сначала эта стратегия успешно работала, поскольку микромиры содержали очень мало объектов, а следовательно, количество возможных действий было невелико и к решению

приводили относительно короткие их последовательности. До того, как была разработана теория вычислительной сложности, господствовало представление, что для больших проблем "масштабирование вверх" было просто вопросом повышения быстродействия оборудования и увеличения объема доступной памяти. Например, оптимизм, сопровождавший разработку систем доказательства теорем, очень быстро улетучился, когда исследователям не удалось добиться доказательства теорем, включающих более нескольких десятков фактов. ➔ *Тот факт, что программа может найти решение в принципе, вовсе не означает, что эта программа уже содержит все механизмы, необходимые для его поиска на практике.*

Иллюзия неограниченности вычислительной мощности касалась не только программ решения задач. Ранние эксперименты в области ▶**машинной эволюции** (теперь она называется **генетическое программирование**) строились на, несомненно, правильном убеждении, что внесение соответствующего ряда небольших изменений в машинный код программы позволит создать эффективную программу решения любой конкретной задачи. Общая идея состояла в том, что необходимо проверять случайные мутации (изменения в коде) с отбором для сохранения мутаций, которые кажутся полезными. Несмотря на тысячи часов потраченного процессорного времени, почти никаких признаков прогресса не было продемонстрировано.

Неспособность справиться с "комбинаторным взрывом" стала одним из основных критических замечаний в адрес ИИ, содержащихся в отчете Лайтхилла ([1412], 1973), что послужило основанием для британского правительства принять решение о прекращении поддержки исследований в области искусственного интеллекта во всех университетах, кроме двух. (Устная традиция рисует в отношении этого решения несколько иную и более красочную картину, с политическими амбициями и личной неприязнью, описание которых выходит за рамки данного курса.)

Третий вид затруднений возник в связи с некоторыми фундаментальными ограничениями базовых структур, использовавшихся в процедурах выработки интеллектуального поведения. Например, Мински и Паперт в книге *Перцептроны* ([1585], 1969) доказали, что перцептроны (простая форма нейронной сети) могут продемонстрировать способность изучить все, что возможно представить с их помощью, но, к сожалению, они позволяют представить лишь очень немногое. В частности, перцептрон с двумя входами не может быть обучен распознавать ситуацию, когда на два его входа поданы разные сигналы. Хотя полученные ими результаты не применимы к более сложным, многослойным сетям, финансирование исследований в области нейронных сетей вскоре сократилось почти до нуля. Как это ни странно, но те новые алгоритмы обучения путем обратного распространения для многослойных сетей, которые стали причиной возрождения необычайного интереса к исследованиям в области нейронных сетей в конце 1980-х годов, фактически впервые были открыты (но в другом контексте) уже в начале 1960-х годов (Келли [1214], 1960; Брайсон [334], 1962).

1.3.4. Экспертные системы (1969–1986)

Основная методология решения проблем, выработанная в течение первого десятилетия исследований в области ИИ, заключалась в использовании поискового механизма общего назначения, посредством которого предпринимались попытки связать воедино элементарные рассуждения так, чтобы отыскать полные решения. Такие подходы получили название ▶**слабые методы** , поскольку, даже будучи достаточно общими, они не масштабируются до уровня больших или сложных проблем. Альтернативой слабых методов является использование более содержательных, предметно-ориентированных знаний, что позволяет строить более длинные цепочки рассуждений и дает возможность проще справиться с теми проблемными ситуациями, которые обычно возникают в специализированных областях знаний. Можно сказать, что для решения сложной проблемы требуется уже почти знать полный ответ.

Одним из первых примеров реализации такого подхода была программа DENDRAL ([339], 1969), разработанная в Стандфордском университете группой ученых, в которую вошли Эд Фейгенбаум (бывший студент Герберта Саймона), Брюс Бьюкенен (философ, который сменил специальность и стал заниматься компьютерными науками) и Джошуа Ледерберг (лауреат Нобелевской премии в области генетики). Эта группа занималась решением проблемы определения структуры молекул на основе информации, полученной от масс-спектрометра. Вход этой программы состоял из химической формулы соединения (например, $C_6H_{13}NO_2$) и спектра масс, позволяющего определять массы различных фрагментов молекулы, который формировался при бомбардировке молекулы потоком электронов. Например, спектр масс может содержать пик в точке $m = 15$, соответствующий массе метилового фрагмента (CH_3).

Первая, простейшая версия этой программы предусматривала генерацию всех возможных структур, совместимых с данной формулой, после чего предсказывала, какой спектр масс должен наблюдаться для каждой из этих структур, сравнивая его с фактическим спектром. Вполне можно ожидать, что такая задача применительно к молекулам более крупных размеров становится практически неразрешимой. Поэтому разработчики программы DENDRAL проконсультировались с химиками-аналитиками и пришли к выводу, что следует попытаться организовать работу по принципу поиска широко известных картин расположения пиков в спектре, которые указывают на наличие общих подструктур в молекуле. Например, для распознавания кетоновых подгрупп (>C=O) с атомной массой 28 имеем следующее.

if M — масса всей молекулы, и есть два пика в x_1 и x_2, таких, что
(a) $x_1 + x_2 = M + 28$; (b) $x_1 - 28$ высокий пик; (c) $x_2 - 28$ высокий пик; и
(d) по крайней мере один из x_1 и x_2 высокий
then это подгруппа кетона.

Выяснение того, что молекула содержит определенную подструктуру, чрезвычайно сокращает число возможных кандидатов. По мнению авторов, программа

DENDRAL стала мощным инструментом, поскольку в нее встроены соответствующие знания, полученные средствами масс-спектроскопии, но не в форме базовых принципов, а в виде эффективных "кулинарных рецептов" (Фейгенбаум и др. [716], 1971). Значение программы DENDRAL определялось тем, что это была первая *успешная* система с интенсивным использованием знаний: ее анализ был построен на использовании большого количества специальных правил. В 1971 году Фейгенбаум и его коллеги в Станфорде приступили к новому проекту эвристического программирования с целью исследования, в какой степени их новая методология построения ▶**экспертных систем** может применяться в других областях.

Следующей крупной попыткой стала разработка системы MYCIN для диагностики инфекции в крови. Исходя из примерно 450 правил, программа MYCIN оказалась способной решать поставленную задачу на уровне некоторых экспертов и заметно лучше молодых врачей. Но у нее было два больших отличия от системы DENDRAL. Во-первых, в отличие от правил DENDRAL, не существует общей теоретической модели, из которой можно было бы вывести правила для MYCIN. Их можно было выделить только из результатов обширного опроса экспертов. Во-вторых, правила должны были отражать неопределенность, связанную с медицинскими знаниями. В MYCIN были включены подпрограммы вычисления неопределенности, называемой ▶**факторами уверенности** (см. главу 13), которые, как казалось на то время, хорошо согласуются с тем, как врачи оценивали влияние отдельных показателей на окончательный диагноз.

Первая успешная коммерческая экспертная система, R1, была развернута в компании DEC (Digital Equipment Corporation) ([1546], 1982). Эта программа помогала составлять конфигурации для выполнения заказов на новые компьютерные системы. К 1986 году она обеспечивала компании DEC экономию примерно 40 миллионов долларов в год. К 1988 году группой искусственного интеллекта компании DEC было развернуто 40 экспертных систем, а в дальнейшем предусматривалось развернуть еще больше. В компании DuPont применялось 100 систем, а в разработке находилось еще 500. Почти в каждой крупной корпорации США была создана собственная группа искусственного интеллекта и экспертные системы либо использовались, либо исследовались.

Важность знания предметной области была также совершенно очевидна в области понимания естественного языка. Несмотря на успех системы SHRDLU Винограда, ее методы нельзя было распространить на более общие задачи. Для решения таких проблем, как разрешение неоднозначности, в ней использовались совсем простые правила, которых, тем не менее, было вполне достаточно для крошечных пределов мира блоков.

Ряд исследователей, в том числе Юджин Чарняк в МТИ и Роджер Шенк в Йельском университете, выдвинули предположение, что уверенное понимание языка потребует общих знаний о мире и некоторого общего метода использования этих знаний. (Шенк пошел дальше, заявив, что "Синтаксиса не существует", и это огорчило многих лингвистов, но послужило началом полезного обсуждения.)

Шенк и его студенты создали серию программ (1978–1981), предназначенных для решения одной и той же задачи — понимания естественного языка. Однако больше внимания при этом уделялось не собственно языку, а проблемам представления знаний и рассуждений, необходимых для его понимания.

Быстро возрастающее количество коммерческих приложений, предназначенных для решения практических задач, потребовало разработки широкого диапазона инструментальных средств представления знаний и рассуждений. Одни из них были основаны на логике, например язык Prolog, ставший популярным в Европе и Японии, или пакет PLANNER в США. В других, следуя идее Мински о ▶**фреймах** ([1582], 1975), был принят более структурированный подход, предполагающий сбор фактов о конкретных типах объектов и событий, с последующим упорядочиванием типов в большую таксономическую иерархию, аналогичную биологической таксономии.

В 1981 году правительство Японии объявило о развертывании проекта "Пятое поколение" — 10-летнего плана по созданию интеллектуальных компьютеров с массовым параллелизмом, работающих под управлением языка Prolog. Бюджет проекта предполагался в размере более 1,3 млрд. долларов по сегодняшнему курсу. В ответ Соединенные Штаты создали корпорацию по микроэлектронике и компьютерным технологиям (MCC) — консорциум, призванный обеспечить стране национальную конкурентоспособность. В обоих случаях область ИИ была частью широкого спектра направлений, включавшего разработку микросхем и исследование интерфейса "человек–машина". В Великобритании Олви в своем отчете потребовал восстановления финансирования тех направлений, финансирование которых было прекращено на основании отчета Лайтхилла. Однако ни один из этих проектов так и не достиг своих амбициозных целей как с точки зрения новых возможностей ИИ, так и с точки зрения экономического эффекта.

В целом в индустрии искусственного интеллекта наблюдался бурный рост, начиная с нескольких миллионов долларов в 1980 году и до миллиарда долларов в 1988 году. Были созданы сотни компаний, разрабатывавших экспертные системы, системы технического видения, роботов, а также специализированное программное обеспечение и аппаратные средства для этих целей.

Однако вскоре после этого наступил период, получивший название "зима искусственного интеллекта", когда многие вновь созданные компании сильно пострадали, поскольку не сумели выполнить свои заманчивые обещания. Оказалось, что создание и поддержка экспертных систем в сложных предметных областях — задача очень сложная, отчасти потому, что используемые в этих системах методы рассуждения отказывали в условиях неопределенности, а отчасти потому, что эти системы были неспособны извлекать уроки из накопленного опыта.

1.3.5. Возвращение к нейронным сетям (1986–настоящее время)

К середине 1980-х годов по меньшей мере четыре разные группы исследователей независимо заново открыли алгоритм обучения путем обратного распростра-

нения, впервые предложенный в начале 1960-х. Этот алгоритм был применен для решения многих проблем обучения в компьютерных науках и психологии, а после публикации результатов его использования в сборнике статей *Распределенная параллельная обработка* ([1933], 1986) привлек всеобщее внимание.

Эти так называемые ▶**коннекционистские** (основанные на соединениях) модели рассматривались многими как непосредственно конкурирующие и с символическими моделями, продвигаемыми Ньюэллом и Саймоном, и с логицистским подходом, предложенным Маккарти и др. Кажется вполне очевидным, что на каком-то уровне мышления люди манипулируют символами. И действительно, антрополог Терренс Дикон в своей книге под названием *Символические виды* ([563], 1997) указал, что это *определяющая характеристика* человека. В противовес этому Джефф Хинтон, ведущий деятель в возрождении нейронных сетей в 1980- и 2010-х годах, назвал символы "светоносным эфиром искусственного интеллекта" — метафорически ссылаясь на несуществующую физическую среду, по которой якобы распространяются электромагнитные волны в представлении физиков начала XIX века. Безусловно, многие концепции, которые можно выделить в языке, при ближайшем рассмотрении не обладают теми или иными логически определенными необходимыми и достаточными состояниями, которые первые исследователи ИИ предполагали сформулировать в аксиоматической форме. Возможно, коннекционистские модели формируют внутренние концепции более гибким и неточным образом, что лучше подходят для беспорядочности реального мира. Они также способны учиться на примерах, сравнивая прогнозируемое ими значение выходного сигнала с истинным значением в решаемой задаче, и изменять свои параметры так, чтобы уменьшить различие, таким образом делая более вероятным получение лучших решений для будущих примеров.

1.3.6. Вероятностные рассуждения и машинное обучение (1987–настоящее время)

Хрупкость создаваемых экспертных систем привела к использованию нового, более широкого научного подхода, включающего вероятности вместо чистой булевой логики, машинное обучение вместо ручного кодирования и экспериментальные результаты вместо философских рассуждений.[14] Теперь обычной практикой стало использование уже существующих теорий вместо предложения совершенно новых, построение вновь выдвигаемых положений на строгих теоремах или твердой экспериментальной методологии, а не на интуиции, и демонстрация приложений, актуальных для реального мира, вместо игрушечных примеров.

[14] Некоторые характеризуют это изменение как победу "чистюль" — тех, кто считает, что теории искусственного интеллекта должны основываться на математической строгости, — над "неряхами" — теми, кто предпочел бы опробовать множество идей, написать несколько программ, а затем оценить то, что кажется работающим. Оба подхода важны. Сдвиг в сторону математической точности подразумевает, что область исследований достигла уровня стабильности и зрелости. Нынешний акцент на глубоком обучении может представлять собой возрождение значимости экспериментаторов.

Совместно используемые эталонные наборы задач стали нормой для демонстрации достигнутого прогресса. Среди них — репозиторий UC Irvine для наборов данных машинного обучения, набор International Planning Competition для проверки алгоритмов планирования, корпус LibriSpeech для задач распознавания речи, набор данных MNIST для распознавания рукописных цифр, наборы ImageNet и COCO для распознавания изображений объектов, набор SQuAD для ответов на вопросы на естественном языке, комплект WMT для машинного перевода и набор International SAT Competitions для проверки выполнимости булевых формул.

В свое время область ИИ появилась отчасти как протест против ограничений в таких уже существовавших областях исследований, как теория управления и статистика. Но в этот период искусственный интеллект включил в себя и положительные результаты из этих областей. Вот что сказал Дэвид Макаллестер ([1523], 1998).

> В ранний период развития ИИ казалось вполне вероятным, что новые формы символических вычислений, например фреймы и семантические сети, сделают основную часть классической теории устаревшей. Это привело к определенной форме самоизоляции, отчего искусственный интеллект в значительной степени отделился от остальной части компьютерных наук. В настоящее время этот изоляционизм преодолен. Появилось признание того, что машинное обучение не следует отделять от теории информации, что проведение рассуждений в условиях неопределенности нельзя изолировать от стохастического моделирования, что поиск не следует рассматривать отдельно от классической оптимизации и управления и что автоматические рассуждения нельзя отделять от формальных методов и статистического анализа.

Эту тенденцию особенно хорошо иллюстрирует область распознавания речи. В 1970-е годы было опробовано большое разнообразие различных архитектур и подходов. Многие из них оказались узко специализированными, надуманными и работали только на нескольких специально отобранных примерах. В 1980-е годы доминирующие позиции в этой области заняли подходы, основанные на использовании ▶**скрытых марковских моделей** (Hidden Markov Model — HMM). Решающую роль здесь сыграли две их особенности. Во-первых, эти модели основаны на строгой математической теории, и это позволило исследователям речи опереться на математические результаты, накопленные в других областях за несколько десятилетий. Во-вторых, они генерируются в процессе обучения программ на крупном массиве реальных речевых данных. Это гарантирует достижение надежных показателей производительности, а в строгих слепых испытаниях модели HMM неизменно улучшают свои показатели. В результате технологии обработки речи и связанная с ними область распознавания рукописных символов смогли совершить переход к широко распространенным промышленным и массовым потребительским приложениям. Обратите внимание, что прежде не было никаких научных утверждений о том, что люди используют HMM для распознавания речи. Эти модели лишь предоставляли математической основу для понимания и решения задач. Однако в разделе 1.3.8 будет показано, что методы глубокого обучения способны в скором времени нарушить эту идиллическую картину.

Год 1988 стал важным годом в отношении установления связей между ИИ и другими научными областями, в том числе статистикой, исследованием операции, теорией принятия решений и теорией управления, методами изучения байесовских сетей и связанных с ними моделей на основе данных. Книга Джуды Перла *Вероятностные рассуждения в интеллектуальных системах* ([1755], 1988) послужила толчком к признанию важности теории вероятностей и теории принятия решений для области искусственного интеллекта. Перл разработал новый подход — модель ▶**байесовских сетей**, обеспечивающей строгий и эффективный формальный подход к представлению нечетко определенных знаний, а также практический алгоритм для вероятностных рассуждений. Эти темы подробно рассматриваются в главах 12–16 вместе с результатами более поздних разработок, позволивших существенно увеличить выразительную мощь вероятностных методов. В главе 20 описываются методы обучения байесовских сетей и связанные с ними модели данных.

Вторым важным итогом 1988 года стала работа Рича Саттона, связавшая метод обучения с подкреплением — этот подход уже использовался в 1950-х годах Артуром Сэмюэлом в его программе игры в шашки — с марковским процессом принятия решений (Markov Decision Processes — MDP), разработанным ранее в рамках теории исследования операций. Последовал целый поток работ по связыванию исследований по планированию ИИ с MDP. С другой стороны, методы обучения с подкреплением нашли применение в области робототехники и управления процессами, одновременно получив глубокое теоретическое обоснование.

Одним из последствий переоценки в области ИИ значения данных, статистического моделирования, оптимизации и машинного обучения стало постепенное воссоединение с такими подобластями, как компьютерное зрение, робототехника, распознавание речи, многоагентные системы и обработка естественного языка, которые к этому времени уже успели несколько отдалиться от основного ядра области искусственного интеллекта. Процесс реинтеграции дал существенное преимущество как в отношении приложений — например, развертывание новых работ по практическому созданию роботов значительно возросло за этот период, — так и в отношении углубления теоретического понимании основных проблем ИИ.

1.3.7. Большие данные (2001–настоящее время)

Замечательные достижения в области увеличения вычислительной мощности и создание Всемирной паутины, World Wide Web, способствовали созданию очень больших наборов данных — явление, сейчас известное, как ▶**большие данные**. Эти наборы данных включают в себя триллионы слов текста, миллиарды изображений и миллиарды часов записей речи и видео, а также колоссальное количество геномных данных, сведений по отслеживанию транспортных средств, информации о посещениях веб-страниц, данных социальных сетей и т.д.

Все это требовало разработки обучающих алгоритмов, специально предназначенных для извлечения преимуществ, даваемых очень большими наборами

данных. Очень часто подавляющее большинство экземпляров данных в таких наборах никак не помечены. Например, в своей известной работе по многозначности слов Яровский ([2400], 1995) указывает, что отдельные вхождения слов, таких как "лист", обычно никак *не маркированы*, чтобы указать, относятся они к флоре, книгоиздательству или материаловедению. Однако при достаточно больших наборах данных подходящие алгоритмы обучения позволяют достичь правильных решений в 96% процентах случаев, исходя из того, какой смысл имеет предложение в целом. Более того, Банко и Брилл ([122], 2001) утверждают, что улучшение в производительности, полученное от увеличения набора данных в размере на два или три порядка по величине перевешивает какие-либо улучшения, которые могут быть получены в результате тонкой настройки алгоритма.

Аналогичный феномен, кажется, имеет место в некоторых задачах компьютерного зрения, таких как заполнение в фотографиях пустых мест, вызванных либо случайными повреждениями, либо такими действиями, как "удаление" бывших друзей. Хейс и Эфрос ([993], 2007) разработали более умный способ достичь этой цели посредством наложения пикселей из аналогичных изображений. Они обнаружили, что этот метод плохо работал на базе данных из нескольких тысяч изображений, но позволял достичь нужного качества при работе с миллионами изображений. Вскоре после появления в базе данных ImageNet десятков миллионов изображений в области компьютерного зрения произошла настоящая революция (Денг и др. [602], 2009).

Доступность больших данных и переход к машинному обучению способствовали восстановлению коммерческого интереса к области искусственного интеллекта (Хавенстейн [986], 2005; Галеви и др. [950], 2009). Большие данные стали решающим фактором, когда в 2011 году система Watson компании IBM добилась победы на людьми — чемпионами викторины Jeopardy! Это событие оказало большое влияние на восприятие искусственного интеллекта широкой публикой.

1.3.8. Глубокое обучение (2011–настоящее время)

Термин ▶**глубокое обучение** относится к машинному обучению с использованием нескольких слоев из простых, регулируемых вычислительных элементов. Эксперименты с такими сетями проводились еще в 1970-х годах, и в форме **сверточных нейронных сетей** они имели некоторый успех в распознавании рукописных цифр в 1990-х годах (ЛеКун и др. [1373], 1995). Однако только в 2011 году методы глубокого обучения действительно получили признание. Сначала это произошло в области распознавания речи, а затем и в области распознавания визуальных объектов.

В 2012 году на конкурсе ImageNet, где требуется классифицировать предлагаемые изображения в одну из тысяч категорий (броненосцы, книжная полка, штопор и т.д.), система глубокого обучения, созданная группой Джеффри Хинтона в Университете в Торонто (Крижевски и др. [1313], 2013) продемонстрировала

значительное улучшение в сравнении с предыдущими системами, построенными главным образом на написанных вручную функциях. С тех пор системы глубокого обучения смогли превзойти возможности человека в решении некоторых визуальных задач (но пока отстают в решении некоторых других задач). Подобные успехи были отмечены и в областях распознавания речи, машинного перевода, постановки медицинского диагноза и компьютерных игр. Использование сети глубокого обучения для представления функции оценки внесло свой вклад в победу программы AlphaGo над ведущими игроками в го со всего мира (Сильвер и др. [2063–2065], 2016–2018).

Эти замечательные успехи привели к возрождению интереса к искусственному интеллекту среди студентов, компаний, инвесторов, правительств, средств массовой информации и широкой общественности. Складывается впечатление, что почти каждую неделю появляются сообщения о том, что очередное новое приложение ИИ смогло приблизиться к человеческим способностям или даже превзойти их. Все это обычно сопровождается очередными спекуляциями либо о безудержном росте достижений, либо о приближении новой зимы ИИ.

Глубокое обучение в значительной степени зависит от мощного оборудования. Если стандартный процессор персонального компьютера способен выполнять от 10^9 до 10^{10} операций в секунду, алгоритм глубокого обучения, работающий на специализированном оборудовании (таком, как GPU, TPU или FPGA), может требовать выполнения от 10^{14} до 10^{17} операций в секунду, в основном в виде сильно распараллеленных матричных или векторных операций. Конечно, глубокое изучение также зависит от доступности в необходимом (большом) количестве данных для обучения, а кроме того, от нескольких алгоритмических трюков (см. главу 21).

1.4. Современное состояние исследований

Проект Стандфордского университета One Hundred Year Study on AI (также известный как AI100) предусматривает регулярный созыв групп экспертов с целью предоставления докладов о текущем состоянии дел в области искусственного интеллекта. В отчете за 2016 год [2338] был сделан вывод, что “можно ожидать существенного увеличения применения приложений ИИ в ближайшем будущем, включая самоуправляемые автомобили, медицинские системы диагностики и таргентной терапии, а также средства оказания физической поддержки в уходе за престарелыми” и что “общество в настоящее время переживает решающий момент в определении того, как развертывать технологии искусственного интеллекта таким образом, чтобы способствовать, а не препятствовать демократическим ценностям, таким как свобода, равенство и прозрачность”. В рамках проекта AI100 на сайте `aiindex.org` также ведется ▶**Индекс ИИ** (AI Index) с целью помочь отслеживать прогресс в этой области. Ниже приведены некоторые основные моменты из отчетов за 2018 и 2019 годы (в сравнении с уровнем 2000 года, принятым как исходный, если не указано иное).

- *Публикации*. Количество публикаций в области ИИ за период с 2010 по 2019 год увеличилось в 20 раз и составило около 20 тысяч в год. Самой популярной категорией было машинное обучение. (На сайте `arXiv.org` в 2009–2017 годы количество статей на тему машинного обучения ежегодно удваивалось.) Следующим по популярности были категории компьютерного зрения и обработки естественных языков.
- *Отношение*. Около 70% новых статей по ИИ имеют нейтральный тон, но количество статей с положительным тоном увеличилось с 12% в 2016 году до 30% в 2018. В большинстве случаев вызывающие беспокойство проблемы носят этический характер: конфиденциальность данных и необъективность алгоритмов.
- *Количество студентов*. В сравнении 2010 годом количество зачисленных на курсы ИИ увеличилось в 5 раз в США и в 16 раз в масштабе всего мира. В настоящее время ИИ — самая популярная специализация в области компьютерных наук.
- *Соотношение полов*. В масштабе всего мира среди профессоров ИИ около 80% мужчин и 20% женщин. Аналогичные цифры имеют место для научных работников, студентов и наемных работников в данной отрасли.
- *Конференции*. С 2012 года посещаемость конференций NeurIPS увеличилась на 800% и достигла уровня 13 500 человек. Для других конференций в области ИИ наблюдается ежегодный рост количества участников — около 30%.
- *Промышленность*. В США количество стартапов в области ИИ выросло в 20 раз, их общее число уже превысило 800.
- *Интернационализация*. В Китае за год публикуется больше статей, чем в США, и примерно столько же, сколько во всей Европе. Однако при оценке публикаций посредством взвешенного показателя цитируемости влияние американских авторов на 50% больше, чем китайских. Сингапур, Бразилия, Австралия, Канада и Индия демонстрируют наиболее быстрый рост количества наемных работников в области ИИ.
- *Компьютерное зрение*. Показатель количества ошибок в распознавании объектов (по результатам конкурса ImageNet LSVRC) улучшился с 28% в 2010 году до 2% в 2017 году, превысив возможности человека. Точность ответов в тестах с открытыми визуальными вопросами (Visual Question Answering — VQA) с 2015 года улучшилась с 55% до 68%, но пока еще отстает от эффективности работы человека — 83%.
- *Скорость работы*. Время обучения для задачи распознавания изображений сократилось в 100 раз за последние два года. Количество вычислительной мощности, используемой в лучших приложениях ИИ, удваивается каждые 3,4 месяца.
- *Обработка языка*. За период с 2015 по 2019 год точность и полнота ответов на вопросы, измеренная в показателях F1 на основании набора данных

Stanford Question Answering Dataset (SQuAD), увеличилась от 60 до 95. На наборе тестовых данных SQuAD2 прогресс оказался больше, с 62 до 90 всего за один год. Оба показателя превышают показатели человеческого уровня.

- *Сравнение с возможностями человека*. Как сообщается, к концу 2019 года системы искусственного интеллекта достигнут или превысят возможности человека по игре в шахматы, го, покер и компьютерную игру PacMan, в викторине Jeopardy! и выполнении тестов ImageNet по распознаванию объектов, в способности распознавания речи в ограниченных областях и перевода с китайского языка на английский в ограниченных областях, в компьютерных играх Quake III, Dota 2, StarCraft II, различных играх игровой приставки Atari, в диагностике рака кожи и простаты, диагностике проблем со сворачиванием белка и диабетической ретинопатии.

Когда (если когда-либо) системы ИИ достигнут производительности на уровне человека при решении широкого круга задач? В проведенных в 2018 году интервью с экспертами в области ИИ был предложен широкий диапазон прогнозируемых дат — от 2029 до 2200 года со средним значением 2099. В аналогичном опросе, проведенном в 2017 году, 50% респондентов считали, что это может произойти к 2066 году, хотя 10% полагали, что это может произойти уже в 2025 году, а некоторые даже ответили “никогда”. Мнения экспертов также разделились в отношении того, потребуются ли нам новые фундаментальные прорывы или будет достаточно лишь простого уточнения и развития уже существующих подходов. Однако не следует принимать эти предсказания слишком серьезно, — как показал в своей работе Филипп Тетлок ([2195], 2017), в области прогнозирования мировых событий эксперты ничем не лучше, чем просто любители.

Как будут работать будущие системы ИИ? Мы пока не можем этого сказать. Как подробно говорилось в этом разделе, в данной научной области за время ее развития сменилось несколько ведущих концепций: в основу была положена смелая мысль, что создать машинный интеллект вообще возможно, при этом принималось, что поставленная цель может быть достигнута путем кодирования экспертных знаний в логические построения. На смену этому пришла уверенность, что главным инструментом могли бы стать вероятностные модели мира, уступившие в последние годы свои позиции убежденности, что машинное обучение позволяет получать такие модели, которые вообще не могут быть созданы на основании любой хорошо изученной теории. Будущее покажет, какая модель будет следующей.

Что ИИ может делать сегодня? Возможно, не так много, как пытались нас убедить авторы некоторых из наиболее оптимистичных статей в СМИ, но все же очень многое. Вот некоторые примеры.

Роботизированные транспортные средства. Отсчет истории роботизированных средств передвижения следует начинать с радиоуправляемых автомобилей 1920-х годов, но первые демонстрации автономного движения по дороге

без специальных направляющих имели место лишь в 1980-х годах (Кенаде и др. [1177], 1986; Дикманс и Зэпп [618], 1987). После успешных демонстраций вождения по грунтовым дорогам на 132-мильном ралли DARPA Grand Challenge в 2005 году (Трун [2212], 2006) и на улицах с дорожным движением в рамках ралли Urban Challenge 2007 года началась настоящая гонка по разработке самоуправляющихся автомобилей. В 2018 году тестовые автомобили компании Waymo преодолели рубеж в 10 миллионов миль езды на дорогах общего пользования без серьезных аварий, при этом человеку-наблюдателю приходилось брать управление на себя только один раз за каждые 6000 миль. Вскоре после этого компания начала предлагать коммерческие услуги роботизированного такси.

В воздухе автономные беспилотники с неподвижным крылом использовались в Руанде для доставки крови через всю страну начиная с 2016 года. Квадрокоптеры автономно выполняли примечательные акробатические маневры, исследуя здания в процессе создания 3-D карт, и самостоятельно собирались в автономные группы.

Перемещение на ногах. Четвероногий робот BigDog, созданный Райбертом и др. ([1846], 2008), перевернул наши представления о том, как роботы могут двигаться, — это уже не походка роботов из голливудских фильмов (медленно, на негнущихся ногах, раскачиваясь из стороны в сторону), а что-то очень похожее на движение животного, способного восстановить равновесие и продолжить движение после толчка или поскользнувшись на замерзшей луже. Атлас, человекоподобный робот, может не только совершать прогулки по неровной местности, но и запрыгивать на ящики и даже делать сальто (Акерман и Джиццо [11], 2016).

Автономное планирование и составление расписаний. Работающая на удалении в сотни миллионов километров от Земли программа Remote Agent агентства NASA стала первой бортовой автономной программой планирования, предназначенной для управления процессами составления расписания операций для космического аппарата (Джонсон и др. [1148], 2000). Программа Remote Agent вырабатывала планы на основе целей высокого уровня, задаваемых с Земли, а также контролировала работу космического аппарата в ходе выполнения этих планов — обнаруживала, диагностировала и устраняла проблемы по мере их возникновения. Сегодня инструментарий планирования EUROPA (Беррейро и др. [133], 2012) используется для планирования рядовых операций марсоходов НАСА, а система SEXTANT (Винтерниц [2364], 2017) обеспечивает автономную навигацию в глубоком космосе, за пределами действия глобальной GPS-системы.

Во время кризиса в Персидском заливе в 1991 году американские войска развернули систему динамического анализа и перепланирования DART (Кросс и Уолкер [503], 1994) с целью автоматизации планирования перевозок и графиков движения для транспорта. Она обрабатывала данные примерно о 50 тысячах автомобилей, единиц грузов и людей одновременно, принимая во внимание их отправные точки, места назначения, маршруты, любые транспортные мощности, возможности портов и аэродромов, разрешая возникающие конфликты по всем параметрам. Управление оборонных проектов США (Defense Advanced Research

Agency Project — DARPA) впоследствии заявило, что одно это приложение с избытком окупило все инвестиции этого ведомства в ИИ за 30 лет.

Ежедневно компании по перевозкам пассажиров и грузов, подобные Uber, и картографические сервисы, такие как Google Maps, обеспечивают эффективный проезд к заданному пункту назначения сотням миллионов пользователей, быстро прокладывая оптимальный маршрут с учетом текущего и *прогнозируемого* трафика по всему пути движения.

Машинный перевод. В настоящее время интерактивные системы машинного перевода позволяют работать с документами более чем на ста языках, в число которых входят родные языки более чем 99% населения Земли. Они переводят сотни миллиардов слов в день для сотен миллионов пользователей. Хотя работа этих систем еще не совсем доскональна, в целом предоставляемый ими перевод вполне адекватен для понимания. Для близкородственных языков (таких, как французский и английский) и при довольно большом объеме обучающих данных перевод в достаточно узкой области приближается к уровню перевода человеком (Ву и др. [2392], 2016).

Распознавание речи. В 2017 году компания Microsoft показала, что ее система распознавания разговорной речи достигла уровня ошибок в словах в пределах 5,1%, что соответствует результатам обычного человека при выполнении тестовой задачи Switchboard, заключающейся в расшифровке телефонных разговоров (Сионг и др. [2394], 2017). Около трети всех компьютерных коммуникаций во всем мире в настоящее время ведется голосовыми данными, а не текстовыми сообщениями, набираемыми с помощью клавиатуры. Приложение Skype обеспечивает перевод речи в режиме реального времени на десять языков. Компании Alexa, Siri, Cortana и Google предлагают программных виртуальных помощников, способных отвечать на вопросы и выполнять поставленные пользователем задания. Например, служба Google Duplex использует средства распознавания и синтеза речи для выполнения бронирования мест в ресторанах по запросу пользователей, обеспечивая поддержку необходимых переговоров от их имени.

Рекомендации, целевая реклама. Такие компании, как Amazon, Facebook, Netflix, Spotify, YouTube, Walmart и другие, используют машинное обучение, чтобы предоставить пользователю информацию о том, что ему может понравиться, исходя из его прошлого опыта и опыта других, подобных ему. Сегмент таких систем имеет долгую историю (Резник и Варян [1874], 1997), но сейчас быстро меняется благодаря новым методам глубокого обучения, обеспечивающим анализ содержания (текста, музыки, видео) наряду с историей поиска и метаданными (Ван ден Орд и др. [2250], 2014; Жанг и др. [2433], 2017). Фильтрацию спама также можно рассматривать как форму рекомендации (или не рекомендации), — современные методы искусственного интеллекта позволяют отфильтровывать более 99,9% спама, а почтовые службы также могут рекомендовать потенциальных получателей наряду с текстом возможного ответа.

Ведение игр. Когда программа Deep Blue компании IBM победила чемпиона мира Гарри Каспарова в матче 1977 года, апологеты человеческого превосходства возложили свои надежды на игру го. Пиет Хут, астрофизик и энтузиаст го, предсказал, что пройдет "сто лет, прежде чем компьютер побьет людей в игре в го, — возможно, даже дольше". Но всего лишь 20 лет спустя программа AlphaGo превзошла всех людей-игроков (Сильвер и др. [2065], 2017). Ке Цзе, чемпион мира, сказал: "В прошлом году программа еще была похожа на человека в своей игре. Но в этом году она предстала как бог го". Программа AlphaGo получила преимущество за счет изучения сотен тысяч игр, сыгранных в прошлом игроками-людьми, а также за счет дистилляции знаний опытных игроков го, работавших в команде.

Следующая программа, AlphaZero, уже не использовала никаких входных данных от человека (за исключением правил игры, конечно же), но оказалась в состоянии на играх с самой собой научиться играть в такой степени, что смогла победить всех противников, и людей, и машин, играя в го, шахматы и сёги (Сильвер и др. [2064], 2018). Как бы там ни было, чемпионы-люди были побеждены системами ИИ в играх столь разнообразных, как викторина Jeopardy! (Ферруччи и др. [736], 2010), покер (Боулинг и др. [272], 2015; Моравчик и др. [1619], 2017; Браун и Сэндхом [321], 2019), а также в видеоиграх DOTA 2 (Фернандес и Молменн [733], 2018), StarCraft II (Винялс и др. [2273], 2019) и Quake III (Ядерберг и др. [1121], 2019).

Распознавание изображений. Не удовольствовавшись превосходством, достигнутым над человеком в точности распознавания объектов в сложном тесте ImageNet, исследователи в области компьютерного зрения переключились на более сложную задачу генерации субтитров к изображению. Вот некоторые впечатляющие примеры: "Человек, едущий на мотоцикле по грунтовой дороге", "Две пиццы на верхней плите кухонной печи" и "Группа молодых людей, играющих в фрисби" (Винялс и др. [2275], 2017). Однако нынешние системы все еще далеки от совершенства: "Холодильник, заполненный большим количеством еды и напитков" оказывается табличкой, запрещающей парковку, частично залепленной множеством маленьких наклеек.

Медицина. Результаты работы алгоритмов ИИ в настоящее время сравнимы или даже превышают результаты врачей — экспертов в диагностике по многим симптомам, особенно когда диагноз ставится на основании изображений. Примеры включают болезнь Альцгеймера (Динг и др. [622], 2018), метастазы рака (Лиу и др. [1433], 2017; Эстева и др. [698], 2017), офтальмологические заболевания (Гульшан и др. [934], 2016) и кожные заболевания (Лиу и др., 2019). Систематический обзор и метаанализ (Лиу и др. [1432], 2019) показали, что производительность программ ИИ в среднем была эквивалентна таковой для специалистов в области здравоохранения. В настоящее время акцент на применении ИИ в медицине часто делается на содействии партнерству между человеком и машиной. Например, система Lyna в диагностике метастатического рака молочной железы достигает общей точности на уровне 99,6% — лучше, чем эксперт-медик без

посторонней помощи, но их комбинация работает еще лучше (Лиу и др. [1434], 2018; Штейнер и др. [2126], 2018).

В настоящее время широкое применение подобных методов ограничивается не их диагностической точностью, а необходимостью продемонстрировать достигнутое улучшение в клинических результатах и необходимостью обеспечения прозрачности, отсутствия предвзятости и сохранения конфиденциальности данных (Топол [2221], 2019). В 2017 году администрация США по продуктам питания и лекарственным средствам (FDA) одобрила только два медицинских приложения ИИ, но в 2018 году их число увеличилось до 12 и продолжает расти.

Климатология. В 2018 году команда ученых получила премию Гордона Белла за создание модели глубокого обучения, способной выявить детальную информацию об экстремальных погодных явлениях, которая ранее была захоронена в огромных массивах климатических данных. Прежде чем программа машинного обучения смогла сделать это, им потребовалось превысить уровень быстродействия в 10^{18} операций в секунду, для чего пришлось использовать суперкомпьютер со специализированным графическим процессором (Курт и др. [1329], 2018). В 2019 году Ролник и его соавторы [1906] представили 60-страничный каталог способов, которыми методы машинного обучения могут быть использованы для отслеживания климатических изменений.

Все сказанное выше — это всего лишь несколько примеров использования систем искусственного интеллекта, которые уже существуют на сегодняшний день. И это не магия или научная фантастика, а наука, инженерия и математика, введением в которые и является эта книга.

1.5. Риски и преимущества искусственного интеллекта

Философ Фрэнсис Бэкон, которому приписывается создание научного метода, в своей книге *Мудрость древних* ([105], 1609) отметил, что “механические искусства можно использовать двояко, они одинаково хороши как для причинения вреда, так и для исцеления”. Поскольку ИИ играет все более важную роль в экономической, социальной, научной, медицинской, финансовой и военной сферах, будет полезно рассмотреть его способности “причинения вреда” и “исцеления”, на современном языке — те риски и преимущества, которые он способен принести. Здесь дается лишь общее представление по этому вопросу, более подробно он анализируется в главах 27 и 28.

Чтобы начать обсуждение преимуществ, сначала просто напомним, что вся наша цивилизация является продуктом интеллекта человека. Если мы получим доступ к существенно большему машинному интеллекту, потолок наших амбиций также существенно поднимется. Потенциал искусственного интеллекта и робототехники в освобождении человечества от монотонной тяжелой работы и резком увеличении производства товаров и услуг может предвещать наступление эры

мира и изобилия. Возможное ускорение научных исследований может иметь следствием, например, нахождение средств излечения различных заболеваний, решения проблемы климатических изменений и дефицита ресурсов. Как сказал Демис Хассабис, генеральный директор компании Google DeepMind, “Сначала решите проблему ИИ, а затем используйте ИИ для решения всех остальных проблем”.

Однако задолго до того, как нам представится возможность “решить проблему ИИ”, мы будем подвержены рискам его неправильного использования, — возможно, непреднамеренного либо наоборот. Одни из этих рисков уже вполне очевидны, тогда как другие пока кажутся лишь весьма вероятными, исходя из существующих тенденций.

- *Смертельное автономное оружие*. По решению Организации Объединенных Наций в эту категорию попадает любое оружие, способное самостоятельно определить местонахождение, выбрать и устранить человеческие цели без вмешательства человека-оператора. Основная проблема с таким оружием заключается в его *масштабируемости*: отсутствие в виде обязательного требования надзора со стороны человека означает, что небольшая группа может развернуть сколь угодно большое количество такого оружия, направив его на людей, определяемых по любому выполнимому этим устройством критерию распознавания. Технологии, необходимые для создания автономного оружия, аналогичны тем, которые требуются для создания автомобилей с автоматическим управлением. В 2014 году в ООН начались неофициальные обсуждения с экспертами потенциальных рисков смертельного автономного оружия, после чего в 2017 году группа правительственных экспертов перешла к официальной стадии, предшествующей заключению соответствующего договора.
- *Наблюдение и убеждение*. В то время как контролировать телефонные линии, каналы видеокамер, электронную почту и другие каналы передачи сообщений людьми дорого, утомительно, а иногда и юридически сомнительно для безопасности персонала, вполне возможно использовать системы ИИ (распознавание речи, компьютерное зрение, обработка естественного языка) в виде масштабируемых приложений для выполнения массового наблюдения за людьми и выявления их деятельности по интересам. Путем адаптации информационных потоков через социальные сети к отдельным лицам на основе использования методов машинного обучения их политическое поведение можно в определенной степени менять и контролировать — проблема, которая стала очевидна на выборах, проходивших в США в 2016 году.
- *Предвзятость принимаемых решений.* Небрежное или преднамеренно неправильное использование алгоритмов машинного обучения для таких задач, как оценка возможности условно-досрочного освобождения или анализ кредитных заявок, может привести к принятию системой ИИ таких решений, которые будут предвзятыми по признаку расы, пола или других

защищенных категорий. Часто данные для обучения сами по себе уже так или иначе отражают господствующие в обществе предубеждения.

- *Влияние на занятость*. Проблема машин, использование которых приводит к сокращению рабочих мест, имеет многовековую историю. История эта никогда не была простой: машины выполняют некоторые из тех операций, которые в противном случае выполняли бы люди, но они также делают людей более продуктивными и, следовательно, дают им больше возможностей для трудоустройства, а компании делают более прибыльными и способными платить более высокую заработную плату. Они также могут сделать некоторые виды деятельности экономически жизнеспособными, тогда как в противном случае они были бы непрактичны. Использование машин в конечном счете приводит к увеличению богатства, но, как правило, способствует смещению этого богатства от труда к капиталу, в еще большей степени усугубляя рост общественного неравенства. Предыдущие достижения в области технологии, такие как изобретение механических ткацких станков, имели следствием серьезные проблемы в области занятости, но в конце концов люди все же находили для себя новые виды работ. С другой стороны, вполне возможно, что со временем системы ИИ будут выполнять и эти новые виды работ. Эта тема быстро становится основным фокусом сосредоточения внимания для экономистов и правительств во всем в мире.
- *Приложения, критические с точки зрения безопасности*. По мере своего развития технологии ИИ все чаще используются в приложениях с "высокими ставками", надежность которых критически важна для безопасности людей, таких как вождение автомобилей или управление водоснабжением городов. Аварии со смертельным исходом уже имели место, и это высветило трудности формальной верификации и статистического анализа риска для систем ИИ, разработанных с использованием методов машинного обучения. Область ИИ действительно нуждается в разработке технических и этических стандартов, по крайней мере сопоставимых с теми, которые приняты в других инженерных и медицинских дисциплинах, где на карту ставится жизнь людей.
- *Кибербезопасность*. Методы искусственного интеллекта полезны для защиты от кибератак, например, путем обнаружения необычных моделей поведения, но они также могут способствовать повышению устойчивости вредоносных программ, повышению их выживаемости и дальнейшему распространению. Например, методы обучения с подкреплением уже были использованы для создания высокоэффективных инструментов проведения автоматизированного персонифицированного шантажа и фишинговых атак.

Более подробно эти темы рассматриваются в разделе 27.3. По мере того как системы ИИ будут становиться все более умелыми, они будут выполнять в обществе все большее количество обязанностей, которые ранее выполняли люди. И подобно

тому как люди, которые ранее выполняли эти обязанности, могли совершать злонамеренные действия, можно ожидать, что в новых условиях люди смогут злонамеренно использовать системы искусственного интеллекта, выполняющие эти обязанности, для причинения еще большего вреда. Все приведенные выше примеры указывают на важность управления и в конечном счете регулирования систем ИИ. В настоящее время исследовательское сообщество и основные корпорации, вовлеченные в исследования в области ИИ, разработали принципы добровольного самоуправления для деятельности, связанной с ИИ (см. раздел 27.3). Правительства и международные организации создают консультативные органы для разработки соответствующих правил для каждого конкретного случая использования ИИ, для подготовки к возможным экономическим и социальным последствиям и для извлечения всех преимуществ из использования возможностей ИИ в решении основных социальных проблем.

Что можно сказать о долгосрочной перспективе? Достигнем ли мы давней цели: создания интеллекта, сравнимого или даже более одаренного, чем человеческий интеллект? И если мы это сделаем, то что тогда?

На протяжении большей части истории искусственного интеллекта эти вопросы находились в тени текущей рутинной работы по созданию систем ИИ, способных сделать что-нибудь, хотя бы отдаленно свидетельствующее о разумности. Как и в любой иной достаточно широкой дисциплине, б*о*льшая часть исследователей в области ИИ специализировалась в определенных направлениях, таких как различные игры, представление знаний, компьютерное зрение или понимание естественного языка, часто исходя из предположения, что прогресс в этим направлении будет способствовать достижению главных целей всей области ИИ. Нильс Нильссон ([1690], 1995), один из первых руководителей проекта Shakey в SRI, напоминал о широте поля этих целей и предупреждал, что каждое из более узких направлений находится в опасности замкнуться на собственных задачах. Позже некоторые влиятельные основатели ИИ, в том числе Джон Маккарти ([1533], 2007), Марвин Мински ([1584], 2007) и Патрик Уинстон (Бил и Уинстон [149], 2009), согласились с предупреждениями Нильссона, предложив исследователям в области ИИ вместо фокусировки на измеримой производительности конкретных приложений обратить взор к истокам и поставить конечной целью своих исследований, как выразился Херб Саймон, создание “машин, которые думают, учатся и создают”. Они назвали эту конечную цель ►**ИИ на уровне человека** (Human-Level AI, **HLAI**) — машина должна научиться делать все, что может сделать человек. Первый симпозиум этого направления состоялся в 2004 году (Мински и др. [1587], 2004). Еще одна попытка создать новое движение в ИИ с аналогичными целями была определена как ►**общий искусственный интеллект** (Artificial General Intelligence — **AGI**) (Гортцел и Пенначин [876], 2007), в его рамках в 2008 году была проведена первая конференция и организован *Journal of Artificial General Intelligence*.

Примерно в то же время были высказаны опасения, что создание ►**искусственного суперинтеллекта** (Artificial SuperIntelligence — **ASI**) — интеллекта,

который намного превосходит человеческие способности — может быть плохой идеей (Юдковски [2416], 2008; Омохундро [1713], 2008). Сам Тьюринг ([2238], 1996) высказал то же самое предположение в лекции, прочитанной им в Манчестере в 1951 году, опираясь на более ранние идеи Сэмюэля Батлера ([353], 1863).[15]

> Кажется вполне вероятным, что после того, как метод машинного мышления будет реализован, ему не понадобится много времени, чтобы превзойти наши слабые силы… Поэтому на каком-то этапе мы должны ожидать, что машины получат контроль — таким способом, который был описан в утопическом романе "Эревон" Сэмюэля Батлера.

В связи с последними достижениями в области глубокого обучения подобные мнения получили еще бо́льшее распространение, о чем свидетельствуют публикации книг, таких как *Суперинтеллект* Ника Бострома ([260], 2014), и публичные заявления Стивена Хокинга, Билла Гейтса, Мартина Риза и Илона Маска.

Вполне естественно испытывать общее чувство беспокойства в отношении идеи создания сверхинтеллектуальных машин. Эту ситуацию можно охарактеризовать как ▶**проблема гориллы**: около семи миллионов лет назад появился вымерший примат, от которого одна ветвь дальнейшего развития ведет к гориллам, а другая — к людям. Сегодня гориллы не слишком счастливы от сосуществования с ветвью человека и по сути не имеют контроля над своим будущим. Если таким же окажется конечный результат достижения успеха в создании сверхчеловеческого ИИ — т.е. люди уступят ему контроль над своим будущим, — то, возможно, мы должны будем прекратить работу над ИИ и как следствие отказаться от выгод, которые он может принести. В этом суть предупреждения Тьюринга: вовсе не очевидно, что мы сможем управлять машинами, которые будут умнее нас.

Если бы сверхчеловеческий ИИ представлял собой черный ящик, который прибыл к нам из космоса, то на самом деле было бы вполне разумно проявлять осторожность в отношении его открытия. Однако это не так: *мы* разрабатываем системы ИИ, поэтому, если в конечном счете они действительно "захватят контроль", как предполагал Тьюринг, то это может быть только результатом неудачного проектирования.

Чтобы избежать такого результата, прежде всего необходимо понять возможные причины потенциальной неудачи. Норберт Винер ([2341], 1960), который задумался над отдаленным будущим искусственного интеллекта, увидев, как программа игры в шашки Артура Сэмюэла научилась обыгрывать своего создателя, сказал по этому поводу следующее.

---

[15] Еще раньше, в 1847 году, Ричард Торнтон, редактор журнала *Primitive Expounder*, протестовал против механических тепломеров: "Ум… обгоняет сам себя и покончил с необходимостью собственного существования, изобретая машины, мыслящие вместо него… Но кто может знать, что такие машины, достигнув большего совершенства, не задумаются о планах исправления всех своих недостатков, а затем, напрягшись, не выдвинут идеи, выходящие за пределы доступного смертному уму!"

> Если, чтобы достичь наших целей, мы используем механическое устройство, в работу которого мы не можем эффективно вмешиваться… нам будет лучше иметь полную уверенность, что цель, поставленная перед машиной, является той, которой мы действительно хотим достичь.

Во многих культурах есть мифы о людях, которые о чем-то просят богов, гениев, магов или дьяволов. В этих историях они неизменно получают то, о чем просили, причем буквально, а затем сожалеют об этом. Третье желание, если оно есть, всегда состоит в том, чтобы отменить первые два. Можно назвать это ▶**проблемой царя Мидаса**: Мидас, легендарный царь в греческой мифологии, попросил богов, чтобы все, к чему он прикоснется, превращалось в золото, но быстро пожалел об этой просьбе, коснувшись своей еды, питья и членов семьи.[16]

Мы упоминали об этой проблеме в разделе 1.1.5, где было указано на необходимость существенной модификации стандартной модели ввода фиксированных целей в машину. Устранение затруднений Винера состоит в том, чтобы вообще не иметь определенной "цели, поставленной перед машиной". Вместо этого нам нужны машины, которые стремятся к достижению целей человека, но знают, что они не имеют полной определенности в том, каковы именно эти цели.

Как ни жаль, почти все исследования ИИ до настоящего времени проводились в рамках стандартной модели, а это означает, что почти весь технический материал в данном издании отражает именно эти интеллектуальные рамки. Есть, однако, некоторые первые результаты и в рамках новых положений. В главе 16 показано, что машина имеет положительный стимул для того, чтобы позволить себе отключиться тогда и только тогда, когда она не уверена в отношении цели человека. В главе 18 были сформулированы и рассмотрены ▶**игры-помощники**, математически описывающие ситуацию, в которой человек имеет цель, и машина пытается ее достичь, но изначально не знает, что она собой представляет. В главе 22 объясняются методы ▶**инвертированного обучения с подкреплением**, позволяющие машинам больше узнать о предпочтениях человека, наблюдая за вариантами выбора, осуществляемого людьми. А в главе 27 исследуются две основные трудности новых концепций: во-первых, наш выбор зависит от наших предпочтений через очень сложную когнитивную архитектуру, которую трудно инвертировать, а во-вторых, мы, люди, можем не ставить постоянные предпочтения на первое место — индивидуально либо как группа, — поэтому может быть неясно, что системы ИИ *должны* делать для нас.

---

[16] Мидас мог поступить лучше, если бы последовал основным принципам безопасности и включил в свою просьбу кнопку "Отменить" и кнопку "Пауза".

Резюме

В данной главе дается определение искусственного интеллекта и описывается исторический контекст, в котором развивалась эта область исследований. Некоторые важные моменты приведены ниже.

- Взгляды ученых на искусственный интеллект могут различаться. Вот два важных вопроса, на которые каждый должен дать ответ: “Вас больше интересует мышление или поведение?” и “Вы намерены моделировать качества человека или просто стремитесь достичь наилучших результатов?”
- В соответствии с тем, что сейчас называют стандартной моделью, ИИ нацелен в основном на **рациональное действие**. Идеальный **интеллектуальный агент** выбирает наилучшее возможное действие в каждой ситуации. В этой книге задача создания интеллектуальных агентов рассматривается именно в этом смысле.
- К этой простой идее необходимо добавить два уточнения. Во-первых, способность любого агента, человека или нет, выбирать рациональные действия ограничена вычислительной сложностью выполнения этого. Во-вторых, концепцию машины, которая преследует определенные цели, следует заменить концепцией машины, преследующей цели, полезные человеку, но не имеющей точного представления о том, что именно они собой представляют.
- Философы (начиная с 400 года до н.э.) обосновали возможность ИИ, выдвинув предположение, что сознание в некотором отношении напоминает машину, оперирующую знаниями, закодированными на каком-то внутреннем языке, и что мышление может использоваться для выбора, какие действия следует предпринять.
- Математики предоставили инструментальные средства для манипулирования высказываниями, обладающими логической достоверностью, а также недостоверными вероятностными высказываниями. Кроме того, они заложили основу не только понимания того, что представляют собой вычисления, но и формирования рассуждений об алгоритмах.
- Экономисты формализовали проблему принятия решений, максимизирующих ожидаемую полезность для лица, принимающего решение.
- Нейробиологи установили некоторые факты о том, как мозг работает и в чем он похож, а чем отличается от компьютеров.
- Психологи подтвердили идею, что люди и животные могут рассматриваться как машины обработки информации. Лингвисты показали, что процессы использования естественного языка вписываются в эту модель.
- Компьютерные инженеры предоставляли все более и более мощные машины, что в конечном счете позволило реализовать самые разные приложения

ИИ, а инженеры-программисты позаботились о том, чтобы они были более удобными для пользователей.

- Теория управления описывает способы разработки таких устройств, которые действуют оптимально на основе обратной связи с окружающей средой. Изначально математические инструментальные средства теории управления весьма отличались от применяемых в области искусственного интеллекта, но эти научные области все больше сближаются.
- История искусственного интеллекта циклична и включает периоды успеха и неоправданного оптимизма, за которыми неизбежно следовали снижение энтузиазма и сокращение финансирования. В ней также были периоды появления новых творческих подходов, лучшие из которых затем неуклонно совершенствовались.
- Область ИИ значительно укрепилась в сравнении с первыми десятилетиями, — как теоретически, так и методологически. Поскольку проблемы, с которыми имел дело ИИ, становились все более сложными, потребовался переход от булевой логики к вероятностной логике, от ручного кодирования знаний к машинному обучению на основе имеющихся данных. Все это уже привело к существенному повышению возможностей реальных систем и более тесной интеграции с другими дисциплинами.
- Поскольку системы ИИ уже нашли себе широкое применение в реальном мире, возникла насущная необходимость рассмотреть широкий диапазон связанных с этим рисков и этических последствий.
- В долгосрочной перспективе мы столкнулись с трудной проблемой контроля над системами искусственного суперинтеллекта, способными развиваться в непредсказуемых направлениях. Решение этой проблемы, похоже, потребует изменения самой концепции искусственного интеллекта.

Библиографические и исторические заметки

Исчерпывающую историю ИИ дал Нильс Нильссон ([1691], 2009), один из пионеров в этой области. Педро Домингос ([633], 2015) и Мелани Митчелл ([1593], 2019) дали обзор машинного обучения, доступный для широкой аудитории, а Кай-Фу Ли ([1376], 2018) описал гонку за международное лидерство в области искусственного интеллекта. Мартин Форд ([757], 2018) взял интервью у 23 ведущих исследователей ИИ.

Основными профессиональными сообществами по ИИ являются Ассоциация развития искусственного интеллекта (Association for the Advancement of Artificial Intelligence — AAAI), группа Special Interest Group in Artificial Intelligence (SIGAI, ранее — SIGART) в рамках Ассоциации вычислительной техники (ACM), Европейская ассоциация искусственного интеллекта и Общество искусственного интеллекта и моделирования поведения (Society for Artificial Intelligence and

Simulation of Behaviour — AISB). Партнерство по ИИ объединяет многие коммерческие и некоммерческие организации, заинтересованные в этических и социальных последствиях внедрения систем ИИ. В журнале *AI Magazine*, выпускаемом AAAI, публикуется множество тематических и учебных статей, а веб-сайт этой ассоциации, `aaai.org`, предлагает посетителям новости, учебные пособия и справочную информацию.

Результаты новейших работ публикуются в трудах основных конференций по ИИ: Международной объединенной конференции по ИИ (International Joint Conference on AI — IJCAI), ежегодной Европейской конференции по ИИ (European Conference on AI — ECAI) и конференции National Conference on AI, чаще упоминаемой под названием AAAI (так сокращенно называется организация American Association for AI, под эгидой которой проводится эта конференция). Машинное обучение освещается в рамках Международной конференции по машинному обучению (International Conference on Machine Learning) и конференции по системам обработки нейронной информации (Neural Information Processing Systems — NeurIPS). Крупнейшими журналами в области общего ИИ являются *Artificial Intelligence*, *Computational Intelligence*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Intelligent Systems* и *Journal of Artificial Intelligence Research*. Имеется также много конференций и журналов, посвященных отдельным областям исследований, которые будут указаны в соответствующих главах.

Упражнения

Эти упражнения предназначены для стимулирования дискуссий, а некоторые могут быть использованы как учебные проекты с установленными сроками. Альтернативный подход — предварительные решения могут быть представлены сейчас, но рассматриваться будут уже после завершения всего курса, построенного на базе этой книги.

1.1. Дайте определения своими словами следующим понятиям: а) интеллект, б) искусственный интеллект, в) агент, г) рациональность, д) логические рассуждения.

1.2. Прочитайте оригинальную статью Тьюринга по искусственному интеллекту (Тьюринг [2235], 1950). В этой статье он обсуждает несколько потенциальных возражений против предложенного им подхода и теста на интеллектуальность. Какие из этих возражений все еще остаются весомыми в определенной степени? Являются ли приведенные им опровержения правильными? Можете ли вы выдвинуть новые возражения, которые следуют из событий, происшедших с тех пор, как Тьюринг написал свою статью? В этой статье он предсказал, что к 2000 году компьютер с вероятностью 30% будет успешно проходить пятиминутный тест Тьюринга с участием слабо подготовленного экспериментатора. Какие шансы, по вашему мнению, имел бы компьютер сегодня? Еще через 50 лет?

1.3. Каждый год премия Лёбнера присуждается программе, наиболее близкой к прохождению определенной версии теста Тьюринга. Проведите исследование и сообщите о последнем лауреате премии Лёбнера. Какие методы используются в этой программе? Какой вклад внесла эта программа в развитие искусственного интеллекта?

1.4. Рациональны ли рефлекторные действия (например, отдергивание руки при прикосновении к горячей печи)? Можно ли считать их интеллектуальными?

1.5. Существуют хорошо известные классы проблем, трудноразрешимых для компьютеров, а также другие классы, для которых доказана их неразрешимость. Означает ли это, что создание искусственного интеллекта невозможно?

1.6. Предположим, мы расширили программу Эванса SYSTEM так, чтобы она могла набрать 200 баллов на стандартном тесте проверки интеллекта (IQ). Можно ли в этом случае считать, что программа стала более интеллектуальной, чем обычный человек? Объясните.

1.7. Нейронная структура морского слизняка аплазии (aplysia) была широко изучена (первым был нобелевский лауреат Эрик Кандел), — просто потому, что у него всего около 20 000 нейронов, большинство из которых крупные и легко доступные для манипуляций. Предположив, что время цикла нейрона аплизии примерно такое же, как и у нейрона человека, сравните вычислительную мощность нейронной структуры животного, выраженную в количестве обновлений памяти в секунду, в сравнении с высокопроизводительным компьютером, данные которого приведены на рис.1.3.

1.8. Почему интроспекция (т.е. самоанализ — составление отчета о собственных мыслях) может оказаться неточной? Может ли человек ошибаться в отношении того, что он думает? Обоснуйте свой ответ.

1.9. В какой степени следующие компьютерные системы можно считать системами искусственного интеллекта:

а) сканеры штрих-кода в супермаркетах;
б) поисковые системы Интернета;
в) голосовое меню телефона;
г) алгоритмы интернет-маршрутизации, динамически реагирующие на состояние сети.

1.10. В какой степени следующие компьютерные системы можно считать системами искусственного интеллекта:

а) сканеры штрих-кода в супермаркетах;
б) голосовое меню телефона;
в) функции проверки правописания и коррекции грамматики в Microsoft Word;
г) алгоритмы интернет-маршрутизации, динамически реагирующие на состояние сети.

1.11. Многие из предложенных вычислительных моделей когнитивной деятельности включают в себя довольно сложные математические операции, такие как свертка изображения с использованием функции Гаусса или поиск минимума функции энтропии. Большинство людей (и конечно, все животные) вообще никогда не изучают подобную математику, почти никто не изучает ее перед поступлением в колледж и почти никто не может вычислить свертку функции с использованием функции Гаусса в уме. Какой смысл в том, чтобы говорить, что "система компьютерного зрения" выполняет подобный тип математических расчетов, тогда как реальный человек не имеет представления о том, как это делается?

1.12. Некоторые авторы утверждают, что самой важной частью интеллекта служат сенсорные способности и моторные навыки, а "высокоуровневые" возможности по сути являются паразитическими, — они просто надстройки над этими основным возможностям. Не вызывает сомнения, что бо́льшая часть эволюции и значительная часть

мозга были связаны с развитием восприятия и моторных навыков, в то время как искусственный интеллект сосредоточился на таких задачах, как ведение игр и формирование логического вывода, которые во многом оказались значительно более простыми по сравнению с восприятием и осуществлением действий в реальном мире. Считаете ли вы, что традиционная направленность искусственного интеллекта на изучение высокоуровневых познавательных способностей является неверной?

1.13. Почему результатом эволюции обычно становится появление систем, которые действуют рационально? Для достижения каких целей предназначены подобные системы?

1.14. Искусственный интеллект — это наука или инженерная дисциплина? Или ни то, ни другое? Поясните свой ответ.

1.15. "Безусловно, компьютеры не могут быть интеллектуальными, ведь они способны выполнять только то, что указали им программисты". Является ли последнее утверждение истинным и следует ли из него первое?

1.16. "Безусловно, животные не могут быть интеллектуальными, ведь они способны выполнять только то, что диктуют им гены". Является ли последнее утверждение истинным и следует ли из него первое?

1.17. "Безусловно, животные, люди и компьютеры не могут быть разумными, ведь они способны выполнять только то, что диктуют законы физики тем атомам, из которых они состоят". Является ли последнее утверждение истинным и следует ли из него первое?

1.18. Изучите литературу по искусственному интеллекту и определите, могут ли в настоящее время компьютеры решать следующие задачи.

а) Игра в настольный теннис (пинг-понг) на достаточно высоком уровне.
б) Вождение автомобиля в центре Каира, Египет.
г) Вождение автомобиля в Викторвилле, Калифорния.
д) Покупка недельного запаса продуктов в супермаркете.
е) Покупка недельного запаса продуктов в интернет-магазине.
ж) Участие в карточной игре бридж на конкурентоспособном уровне.
з) Открытие и доказательство новых математических теорем.
и) Написание рассказа, который непременно должен быть смешным.
к) Предоставление компетентных юридических консультаций в специализированной области права.
л) Перевод разговорной речи в режиме реального времени с английского языка на шведский.
м) Выполнение сложной хирургической операции.

1.19. В отношении тех задач из упражнения 1.18, которые в настоящее время являются неосуществимыми, попытайтесь понять, в чем заключаются основные трудности, и предсказать, когда они будут преодолены (и произойдет ли это вообще).

1.20. Уже давно в различных подобластях ИИ проводятся конкурсы в виде постановки стандартной задачи и предложения исследователям сделать все возможное для нахождения наилучшего ее решения. Примерами могут служить соревнование DARPA Grand Challenge для роботизированных автомобилей, международный турнир по планированию, роботизированная футбольная лига Robocup, конкурс TREC по извлечению информации и соревнования по машинному переводу и распознаванию речи. Изучите материалы пяти из этих конкурсов и опишите прогресс, достигнутый за последние годы. В какой степени эти конкурсы способствовали продвижению к современному состоянию исследований в области ИИ? В какой степени они наносят ущерб этой области, отнимая возможности реализации у новых идей?