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