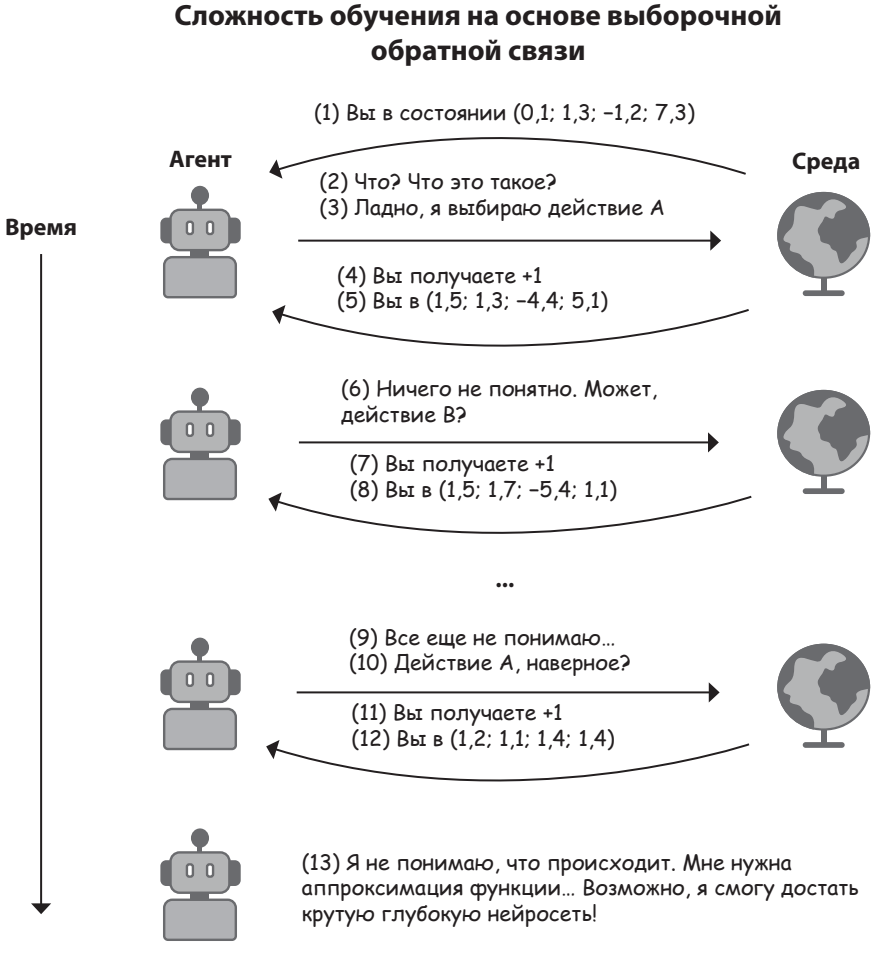
В главе 4 мы отдельно изучим все тонкости оценочной обратной связи. То есть ваши программы будут обучаться на одновременно одинарной (в отличие от последовательной), оценочной и исчерпывающей (в отличие от выборочной) обратной связи.

Агенты глубокого обучения с подкреплением учатся на выборочной обратной связи

Получаемая агентом награда — просто образец. В действительности у агента нет доступа к функции вознаграждения. К тому же состояние и пространство действий обычно довольно большие или даже бесконечные, что затрудняет обучение с использованием рассеянной и слабой обратной связи. Поэтому агент должен быть способен обобщать и учиться на выборочной обратной связи.

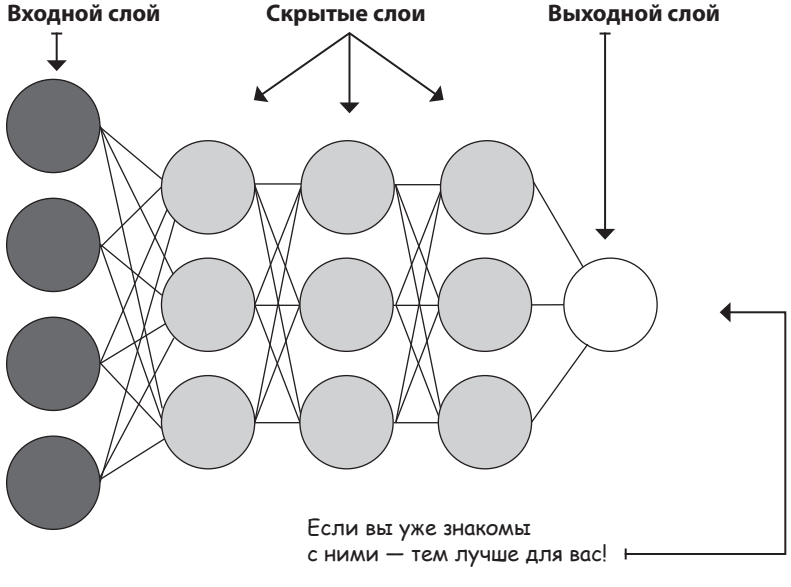


Агенты для аппроксимации правил называются *ориентированными на правила*, для аппроксимации функций ценности — *ценностно ориентированными*, для аппроксимации моделей — *модельно-ориентированными*, а агенты для аппроксимации и правил, и функций ценности называются *«актеры-критики»*. Агенты могут предназначаться для аппроксимации одного из этих компонентов или сразу нескольких.

Агенты глубокого обучения с подкреплением используют мощную аппроксимацию нелинейных функций

Агент может аппроксимировать функции с помощью разных методов и подходов, от деревьев принятия решений до SVM и нейросетей. Но в этой книге мы ограничимся только последними. В конце концов, именно они делают RL глубоким. Такое решение подходит не для всех задач: нейросети требовательны к данным и сложны для интерпретации — помните об этом. Но на сегодня это один из самых действенных способов аппроксимации функций, который часто показывает непревзойденную производительность.

Простая нейросеть с прямой связью



Искусственная нейронная сеть (ИНС) — это многоуровневый аппроксиматор нелинейных функций, отдаленно напоминающий биологические нейросети в мозге животного. ИНС — это не алгоритм, а структура, состоящая из нескольких слоев математических преобразований, применяемых к входным значениям.

Главы 3–7 посвящены только задачам, в которых агенты обучаются на исчерпывающей (а не выборочной) обратной связи. В главе 8 мы впервые рассмо-

трим полную задачу DRL: использование нейросетей для обучения агента на выборочной обратной связи. Помните, что связь, на которой обучаются агенты DRL, одновременно последовательная, оценочная и выборочная.

Прошлое, настоящее и будущее глубокого обучения с подкреплением

Для приобретения навыков не обязательно углубляться в историю, но, зная ее, вы сможете лучше вникнуть в контекст изучаемой темы. Это может повысить вашу мотивацию и улучшить ваши навыки. Ознакомившись с историей ИИ и DRL, вы поймете, чего можно ожидать от этой перспективной технологии в будущем. Иногда мне кажется, что такое количество внимания ИИ идет только на пользу, привлекая людей. Но, когда пора приниматься за работу, ажиотаж утихает, и это проблема. Я не против того, чтобы люди восторгались ИИ, но мне хочется, чтобы их ожидания были реалистичными.

Новейшая история искусственного интеллекта и глубокого обучения с подкреплением

История DRL началась очень давно. Еще в древности люди задумывались о возможности существования разумных созданий, помимо людей. Но отправной точкой можно считать работы Алана Тьюринга (Alan Turing) в 1930–1950 годах, проложившие путь к современной информатике и ИИ и послужившие основой для последующих научных изысканий в этой области.

Самый известный пример его трудов — тест, который предлагает стандартный подход к оценке компьютерного интеллекта: если в ходе сеанса вопросов/ответов наблюдателю не удается отличить компьютер от человека, первый считается разумным. Несмотря на свою примитивность, тест Тьюринга позволил целым поколениям размышлять о возможности создания разумных машин, определив цель, на которую могут ориентироваться исследователи.

Формальное начало ИИ как академической дисциплины можно отнести к Джону Маккарти (John McCarthy), влиятельному исследователю ИИ, который внес заметный вклад в эту область. В 1955 году Маккарти впервые предложил термин «искусственный интеллект», в 1956-м — возглавил конференцию по ИИ, в 1958-м — изобрел язык программирования Lisp, а в 1959-м — стал соучредителем лаборатории MIT, которая занимается исследованием ИИ. Несколько десятилетий он публиковал важные научные работы, способствовавшие развитию ИИ как области научных исследований.

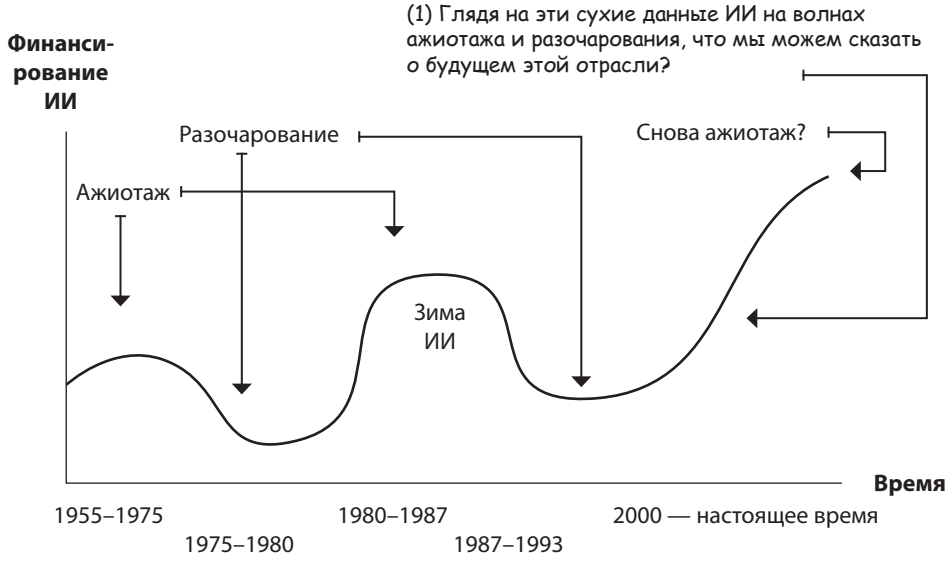
Зимы искусственного интеллекта

Вся та работа и прогресс, которые наблюдались на ранних этапах развития ИИ, вызывали большой интерес, но не обошлось и без серьезных неудач. Известные исследователи высказывались о том, что человекоподобный компьютерный

интеллект появится в течение нескольких лет, но этого так и не произошло. Ситуация усугубилась, когда знаменитый ученый Джеймс Лайтхилл (James Lighthill) составил отчет, в котором раскритиковал положение дел в академическом исследовании ИИ. Все это привело к началу длинного периода, на протяжении которого исследования в этой области испытывали сокращение финансирования и общественного интереса, — к *зиме искусственного интеллекта*.

На протяжении многих лет в области изучения ИИ сохранялась такая картина: исследователи добивались успехов, что порождало чрезмерный оптимизм и завышенные ожидания со стороны общества и в итоге приводило к сокращению финансирования от правительства и отраслевых партнеров.

Модель финансирования ИИ на протяжении многих лет



Текущее положение дел в сфере искусственного интеллекта

Скорее всего, мы переживаем еще один весьма оптимистичный период в истории ИИ, поэтому должны сохранять бдительность. Те, кто применяет ИИ на практике, понимают, что это мощный инструмент, но некоторые люди считают ИИ волшебным ящиком, который может принять любую задачу и выдать лучшее возможное решение. Это не так. Кто-то даже выражает реальные опасения по поводу того, что у ИИ может пробудиться сознание. Вот что сказал Эдсгер В. Дейкстра (Edsger W. Dijkstra) по этому поводу: «Вопрос о том, может ли компьютер мыслить, ничуть не интересней вопроса о том, может ли подводная лодка плавать».

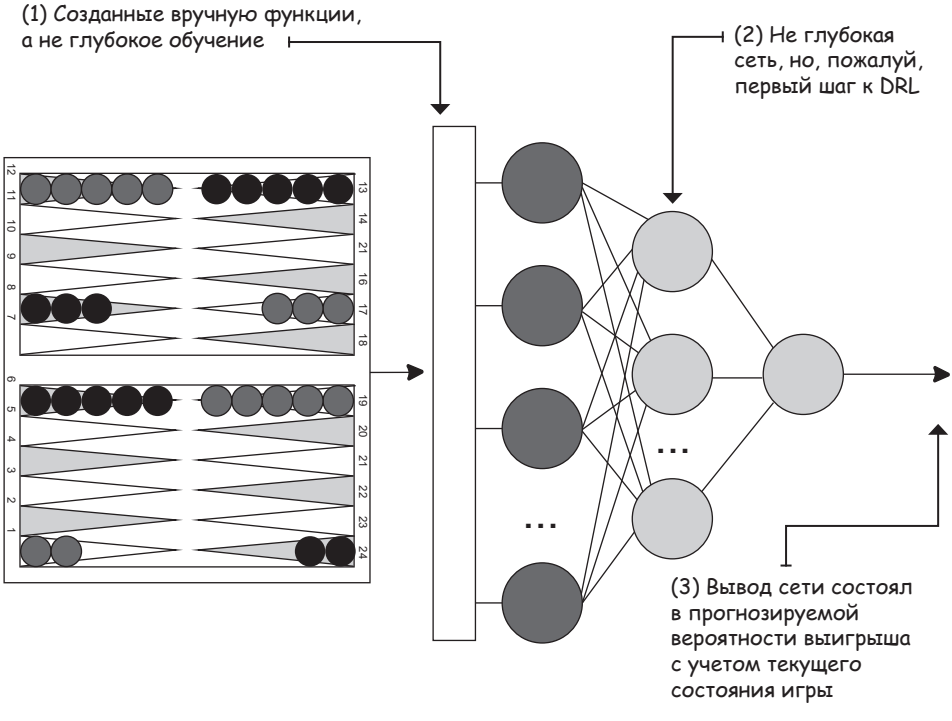
Но если не брать во внимание эту «голливудскую» версию ИИ, последний прогресс в этой области вселяет оптимизм. На сегодня самые влиятельные

компании в мире делают крупные инвестиции в исследование искусственного интеллекта. Google, Facebook, Microsoft, Amazon и Apple вкладывают много средств в эту область и отчасти обязаны ей своей высокой прибыльностью. Существенные и стабильные инвестиции создали идеальные условия для текущих темпов исследования ИИ. Ученые имеют доступ к самым мощным компьютерам и огромным объемам данных. Команды ведущих исследователей работают совместно над одними и теми же задачами. Область ИИ стала более стабильной и продуктивной. Мы наблюдаем один успех за другим, и в ближайшем будущем эта тенденция лишь усилится.

Прогресс глубокого обучения с подкреплением

Искусственные нейросети начали применять для выполнения задач RL в 1990-х. Классический пример — программа для игры в нарды, TD-Gammon, созданная Джеральдом Тезауро (Gerald Tesauro) и др. Чтобы освоить нарды, она сначала научилась самостоятельно оценивать позиции на доске с помощью RL. И хотя методики, реализованные в TD-Gammon, — это не совсем DRL, программа стала одним из первых случаев успешного применения ИНС для решения сложных задач RL.

Структура TD-Gammon



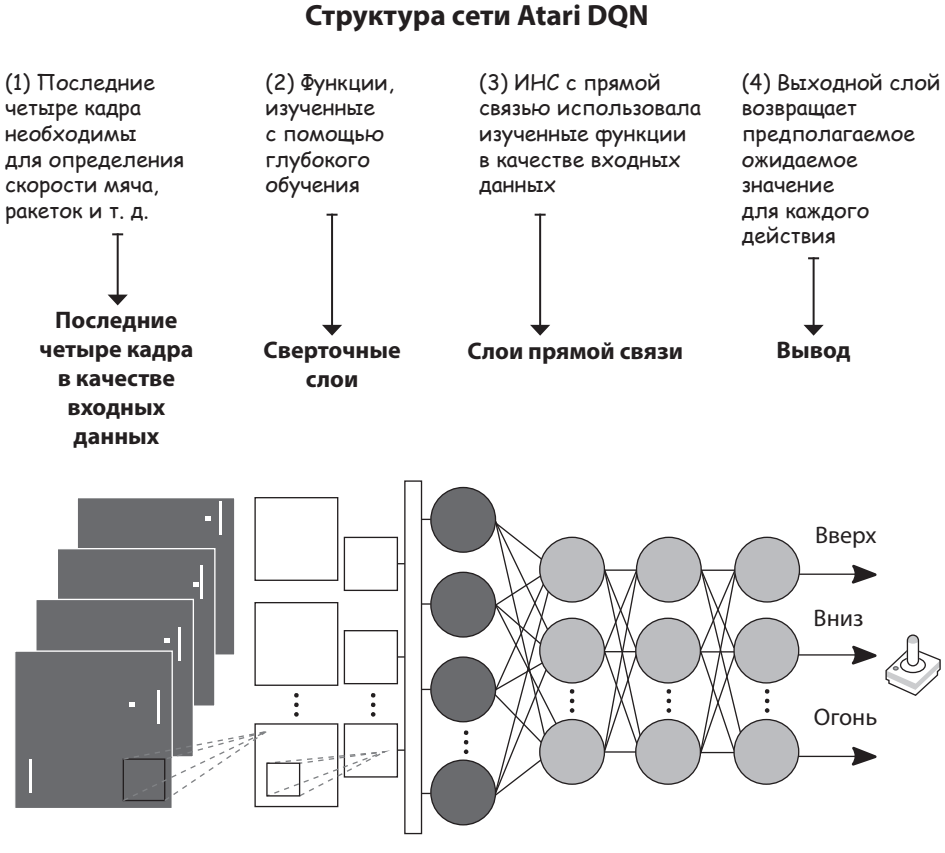
В 2004 году Эндрю Ын (Andrew Ng) и др. разработали автономный вертолет, который самостоятельно обучался трюкам высшего пилотажа, часами наблюдая за полетами опытных пилотов. При разработке была применена методика *обратного обучения с подкреплением*, при которой агент учится на демонстрациях специалистов. В том же году Нейт Кохл (Nate Kohl) и Питер Стоун (Peter Stone) применили категорию методов DRL, известную как *градиент политик*, чтобы создать робота, который играет в футбол, для турнира RoboCup. Для обучения агента движению вперед ученые использовали RL. Спустя всего три часа этот робот научился двигаться вперед быстрее, чем любой другой с той же аппаратной начинкой.

В 2000-х были и другие успехи, но по-настоящему область DRL начала развиваться только с 2010 года, когда произошел всплеск популярности глубокого обучения. В 2013 и 2015 годах Мних (Mnih) и др. опубликовали несколько научных работ с описанием алгоритма DQN (Deep Q-Learning), который учился играть в приставку Atari по одним лишь пикселям на экране. Используя сверточную нейросеть (convolutional neural network, CNN) и единый набор гиперпараметров, алгоритм DQN превзошел профессиональных игроков в 22 играх из 49.

Это достижение положило начало революции в сообществе DRL: в 2014 году Сильвер (Silver) и др. выпустили алгоритм градиента по детерминированным политикам (deterministic policy gradient, DPG), а уже через год Лилликрап (Lillicrap) и др. представили его улучшенную, глубокую версию (deep deterministic policy gradient, DDPG). В 2016 году Шульман (Schulman) и др. предложили методы оптимизации стратегий в доверительной области (trust region policy optimization, TRPO) и обобщенной оценки преимущества (generalized advantage estimation, GAE). Сергей Левин (Sergey Levine) и др. опубликовали управляемый поиск политик[1] (Guided Policy Search, GPS), а Сильвер и др. показали AlphaGo в этом же году и AlphaZero в следующем. В этот период появилось много других алгоритмов: двойные глубокие Q-сети (double deep Q-networks, DDQN), приоритетное воспроизведение опыта (prioritized experience replay, PER), оптимизация проксимальной политики (proximal policy optimization, PPO), «актер-критик» с воспроизведением опыта (actor-critic with experience replay, ACER), асинхронное преимущество «актер-критик» (asynchronous advantage actor-critic, A3C), преимущество «актер-критик» (advantage actor-critic, A2C), «актер-критик» с использованием области доверия с коэффициентом Кронекера (actor-critic using Kronecker-factored trust region, ACKTR), Rainbow («Радуга»), Unicorn («Единорог»)

[1] В книге встречаются термины policy и strategy. Чтобы не возникало путаницы, первый обозначен как политика, второй — стратегия. — *Примеч. пер.*

(это, кстати, настоящие названия) и т. д. В 2019 году Ориол Виньялс (Oriol Vinyals) и др. показали агент AlphaStar, способный обыгрывать профессиональных игроков в StarCraft II. Спустя несколько месяцев Якуб Пахоцки (Jakub Pachocki) и др. наблюдали за тем, как их команда ботов для игры в Dota 2 под названием Five стала первым ИИ, победившим чемпионов мира по киберспорту.



Благодаря прогрессу DRL за последние два десятилетия мы прошли путь от нард с 10^{20} состояниями полной информации до игры го или, что еще лучше, StarCraft II, в которых таких состояний 10^{170} и 10^{270} соответственно. Сложно представить более удачный момент для того, чтобы начать знакомство с этой областью. Подумайте только, что может произойти за следующие 20 лет! Хотите быть причастны к этому? Область DRL сейчас на подъеме, и я ожидаю, что она продолжит бурно развиваться.