грокаем

Глубокое обучение

с подкреплением

Мигель Моралес

ББК 32.813+32.973.23-018
УДК 004.89+004.85
М79

Моралес Мигель

М79 Грокаем глубокое обучение с подкреплением. — СПб.: Питер, 2023. — 464 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-4461-3944-6

Мы учимся, взаимодействуя с окружающей средой, и получаемые вознаграждения и наказания определяют наше поведение в будущем. Глубокое обучение с подкреплением привносит этот естественный процесс в искусственный интеллект и предполагает анализ результатов для выявления наиболее эффективных путей движения вперед. Агенты глубокого обучения с подкреплением могут способствовать успеху маркетинговых кампаний, прогнозировать рост акций и побеждать гроссмейстеров в Го и шахматах.

Давайте научимся создавать системы глубокого обучения на примере увлекательных упражнений, сопровождаемых кодом на Python с подробными комментариями и понятными объяснениями. Вы увидите, как работают алгоритмы, и научитесь создавать собственных агентов глубокого обучения с подкреплением, используя оценочную обратную связь.

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.813+32.973.23-018
УДК 004.89+004.85

ISBN 978-1617295454 англ.
ISBN 978-5-4461-3944-6

Оглавление

Введение в глубокое обучение с подкреплением

1

В этой главе

- ✓ Вы узнаете, что такое глубокое обучение с подкреплением и чем оно отличается от других подходов к машинному обучению.
- ✓ Познакомитесь с последними достижениями в сфере глубокого RL и узнаете, как оно может помочь в решении разных задач.
- ✓ Узнаете, чего ожидать от этой книги и как извлечь из нее максимум пользы.

Я представляю время, когда мы будем для роботов тем же, чем сейчас собаки являются для людей, и болею за машины.

Клод Шеннон, отец информационного века, выдающийся ученый в области искусственного интеллекта

Люди хотят быть счастливыми. Каждое наше действие, от выбора еды на завтрак до продвижения по карьерной лестнице, обусловлено стремлением привнести в жизнь приятные моменты. Это могут быть эгоцентричные удовольствия или благородные цели, то, что приносит немедленное или долгосрочное удовлетворение. Главное, что мы считаем это важным и ценным, ведь в каком-то смысле эти мгновения придают нашей жизни смысл.

Наша способность испытывать чувства от подобных моментов связана с интеллектом — способностью приобретать и применять знания и навыки. Люди, которых общество считает умными, могут отказываться не только от немедленного удовлетворения в пользу долгосрочных целей, но и от хорошего, гарантированного будущего в пользу лучшего, но неопределенного. Целей, которые дольше материализуются и обладают неизвестной долгосрочной ценностью, обычно сложнее всего достичь. Преодолеть все эти трудности могут исключительные люди — интеллектуалы и лидеры, признанные в обществе.

Из книги вы узнаете о подходе глубокого обучения с подкреплением, связанном с созданием компьютерных программ, способных достигать целей, требующих интеллекта. В первой главе вы познакомитесь с основами этого подхода и получите советы о том, как извлечь из моего пособия максимум пользы.

Что такое глубокое обучение с подкреплением

Глубокое обучение с подкреплением (deep reinforcement learning, DRL) — это подход к искусственному интеллекту на основе машинного обучения, направленный на создание компьютерных программ, способных выполнять задачи, требующие интеллекта. Отличительная черта программ DRL — обучение методом проб и ошибок на основе обратной связи, которая является одновременно последовательной, оценочной и выборочной за счет использования мощной аппроксимации нелинейных функций.

Давайте разберем это определение по частям, но не слишком увлекайтесь деталями, ведь у вас впереди еще целая книга, чтобы погрузиться в глубокое обучение с подкреплением. Эта глава — введение в материал, который вы будете изучать далее. Периодически мы будем возвращаться к ней при рассмотрении подробностей в последующих главах.

Цель этой книги — дать вам полное, всестороннее понимание этого определения. После прочтения вы сможете объяснить, почему я выбрал именно эти слова и именно эти формулировки, а пока просто расслабьтесь и прочтите первую главу.

Глубокое обучение с подкреплением — это подход к искусственному интеллекту на основе машинного обучения

Искусственный интеллект (ИИ) — это раздел информатики, связанный с созданием программ, способных демонстрировать разумное поведение. Традиционно любое программное обеспечение (ПО), отображающее такие когнитивные способности, как восприятие, поиск, планирование и обучение, считается частью ИИ. Вот несколько примеров:

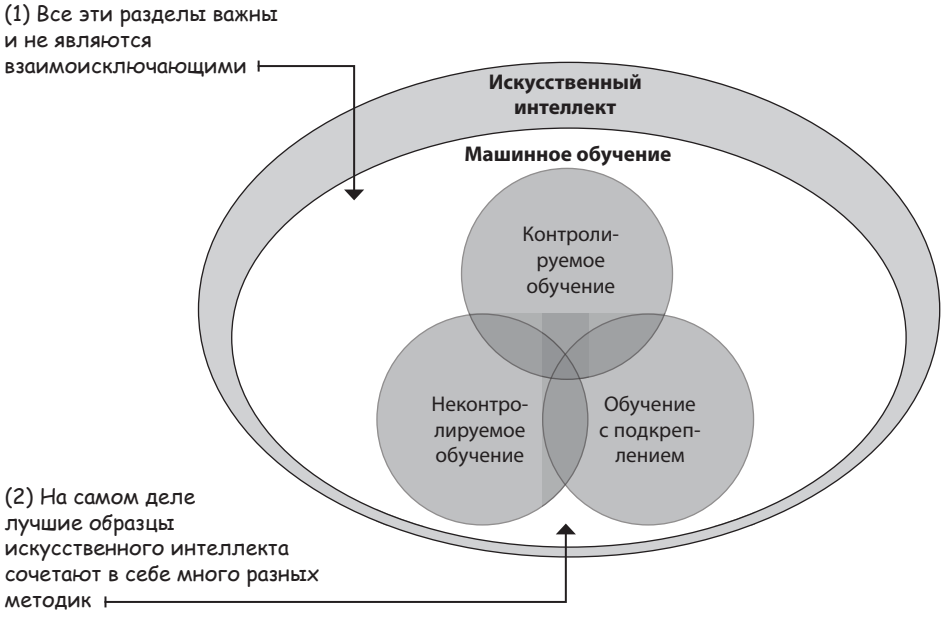
- страницы, возвращаемые поисковой системой;
- маршрут, прокладываемый GPS-навигатором;
- умный помощник с распознаванием голоса и синтетической речью;
- список рекомендаций на сайтах интернет-магазинов;
- функция «следуй за мной» в дронах.

Области искусственного интеллекта

Любая программа, проявляющая интеллект, относится к ИИ, но не все примеры искусственного интеллекта могут обучаться. *Машинное обучение* — это область ИИ, посвященная созданию ПО для выполнения задач, требующих

интеллекта, через обучение на основе данных. У ML есть три основных направления: контролируемое, неконтролируемое и обучение с подкреплением.

Основные направления машинного обучения



Контролируемое обучение (supervised learning, SL) предполагает использование промаркированных данных. В процессе SL человек решает, какие данные нужно собрать и как их пометить. Цель этого направления ML — обобщение. Классический пример — приложение для распознавания цифр, написанных от руки: человек собирает изображения с рукописными цифрами и учит модель правильно распознавать и категоризировать эти цифры. Ожидается, что обученная модель сможет обобщать и категоризировать новые изображения с такими цифрами.

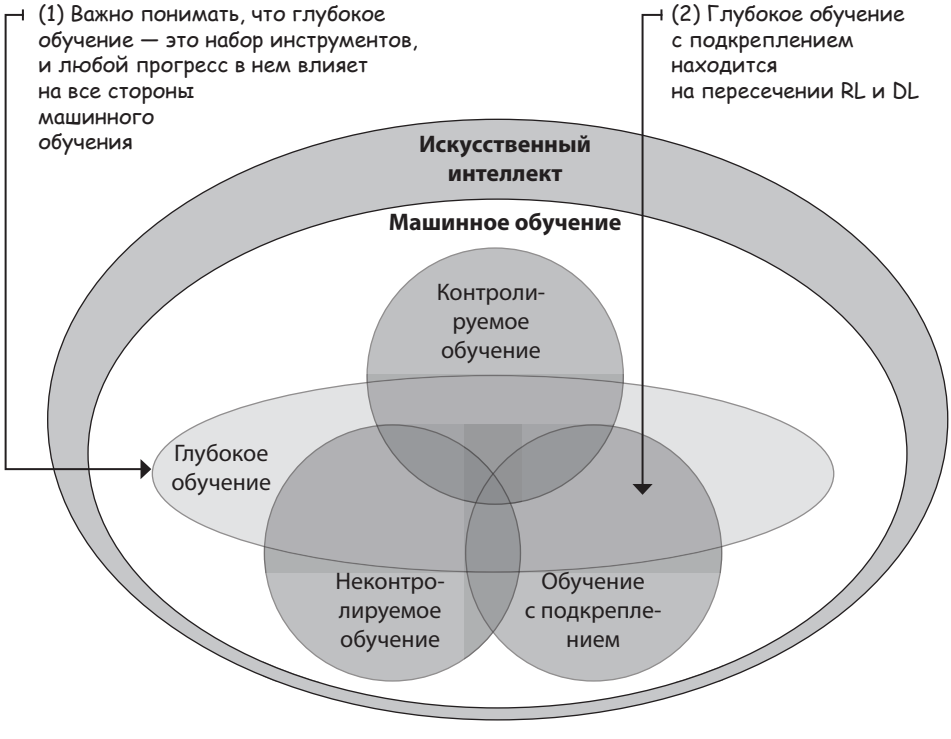
Неконтролируемое обучение (unsupervised learning, UL) подразумевает использование непромаркированных данных. Несмотря на то что данные больше не нуждаются в метках, методы по их сбору, которые использует компьютер, все еще должны разрабатываться человеком. Цель UL — сжатие. Классический пример — приложение для сегментации клиентов: человек собирает сведения о клиентах и учит модель объединять их в кластеры, которые сжимают информацию, раскрывая базовые закономерности.

Обучение с подкреплением проводится методом проб и ошибок. В задачах этого типа человек не маркирует данные, не собирает их и не участвует в разработке механизма их сбора. Цель RL — действие. Классический пример — агент

для игры в Pong, который взаимодействует с эмулятором аркадного автомата Pong и учится, выполняя действия и наблюдая за их последствиями. Обученный агент должен уметь действовать таким образом, который позволит ему успешно играть в Pong.

Относительно новый действенный подход к ML, *глубокое обучение*, включает использование многоуровневой аппроксимации нелинейных функций, обычно в виде нейронных сетей. DL не самостоятельный раздел ML, поэтому принципиально не отличается от описанных выше методов. Это набор техник и методов использования нейронных сетей для выполнения задач ML: будь то SL, UL или RL. DRL — это всего лишь подход к решению задач RL с использованием DL.

Глубокое обучение — это мощный инструментарий



Суть в том, что DRL — это подход к решению задачи. Область ИИ определяет следующую задачу: создание разумных машин. Один из способов ее решения — DRL. На страницах книги вы встретите сравнения между RL и другими видами машинного обучения, но в этой главе приводятся только определения и исторический экскурс в ИИ в целом. Важно отметить, что область DRL — это часть RL, поэтому при упоминании RL я имею в виду и DRL. Но при необходимости я провожу между ними различие.

Глубокое обучение с подкреплением предназначено для создания компьютерных программ

По сути, в DRL мы занимаемся сложными последовательными задачами принятия решений в условиях неопределенности. Но эта тема исследуется и в других областях. Например, *теория управления* (control theory, CT) изучает пути управления сложными известными динамическими системами. В CT динамика систем, которыми мы пытаемся управлять, известна заранее. *Исследование операций* (operations research, OR) тоже посвящено принятию решений в условиях неопределенности. Но задачи в этой области обычно гораздо шире, чем в DRL. *Психология* изучает человеческое поведение. Это отчасти та же «сложная последовательная задача о принятии решений в условиях неопределенности».

Синергия между схожими направлениями

Подытожим: вы имеете дело с областью, на развитие которой влияет много других направлений. Это здорово, но возможна несогласованность терминологий, обозначений и т. д. Я предпочитаю подходить к этой проблеме в кон-

тексте компьютерных наук, поэтому посвятил свою книгу созданию программ, решающих сложные задачи принятия решений в условиях неопределенности. В связи с этим на ее страницах встречаются примеры кода.

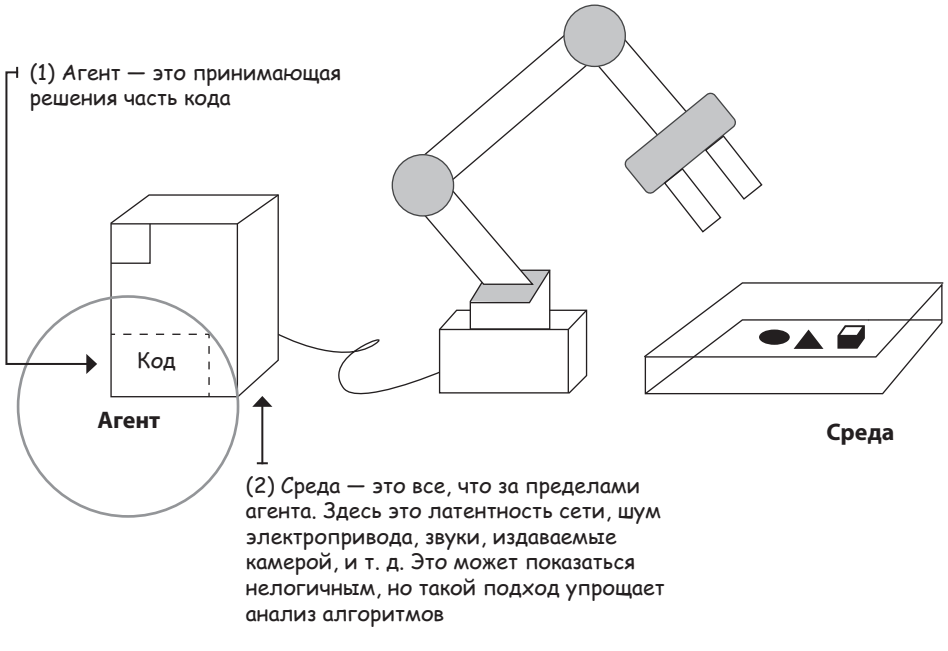
В DRL такие программы называются *агентами*. Они занимаются только принятием решений. Это значит, что рука робота, обученного поднимать объекты, не входит в состав агента. Агент — это только код, принимающий решения.

Агенты глубокого обучения с подкреплением могут выполнять задачи, требующие наличия интеллекта

Противоположность агента — *среда* — все, что находится за пределами агента и вне его полного контроля. Давайте снова представим себе робота, которого вы учите поднимать предметы. Объекты, которые нужно поднять, поднос, на котором они лежат, ветер и все остальное, что не принадлежит лицу, принимающему решения, — это часть среды. Значит, рука робота тоже относится к среде, так как не входит в состав агента. Агент может принять решение о движении рукой, но само движение создает шум, поэтому рука — это часть среды.

Поначалу это четкое разделение между агентом и средой может показаться нелогичным, но агенту отводится лишь одна роль: принятие решений. Все, что происходит после, относится к среде.

Граница между агентом и средой

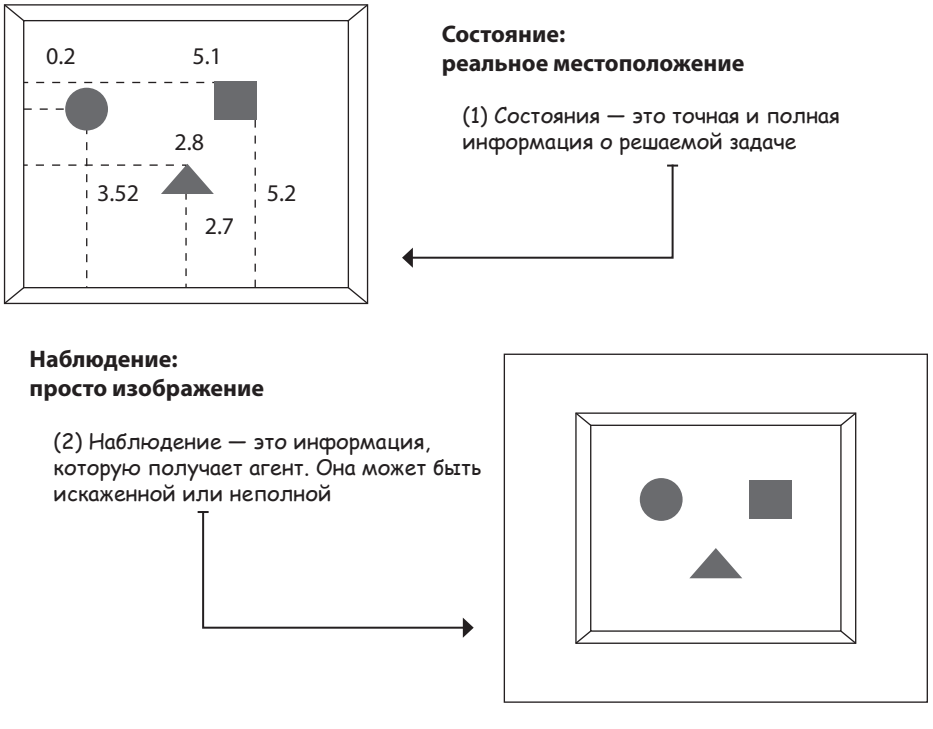


Ниже приведен общий обзор всех компонентов DRL. Подробнее о них вы узнаете в главе 2.

Среда — это набор переменных, относящихся к задаче. Например, в случае с ранее упомянутым роботом в ее состав входят такие переменные, как местоположение, скорость и направление движения руки. Переменные и все их возможные значения называют *пространством состояний*, где состояние — это отдельный экземпляр, набор значений, которые принимают переменные.

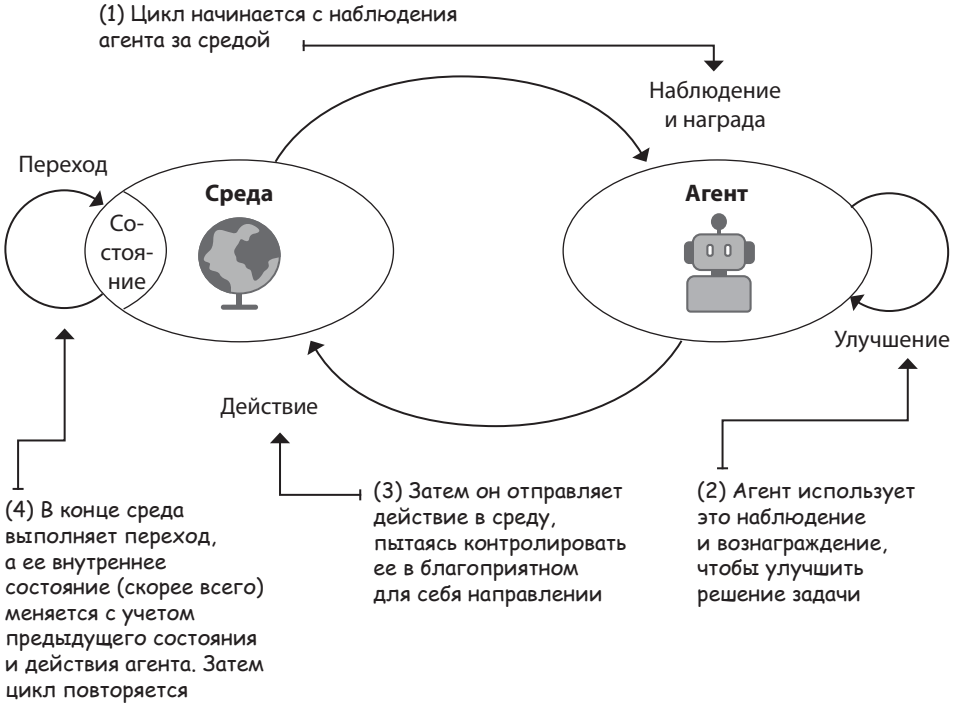
Интересно, что у агентов обычно нет доступа ко всему состоянию среды. Та часть состояния, которую агент может наблюдать, называется *наблюдением*. Наблюдения зависят от состояний, но представляют собой то, что может видеть агент. Например, в случае с роботизированной рукой агент может иметь доступ только к изображениям из камеры. У каждого объекта есть определенное местоположение, но агенту ничего не известно об этой конкретной части состояния. Вместо этого он воспринимает основанные на состояниях наблюдения. В научной литературе, в том числе и в этой книге, наблюдения и состояния часто используются как синонимы. Заранее прошу прощения за такую непоследовательность. Просто помните о различиях и делайте поправку на лексику — это главное.

Состояния и наблюдения



В каждом состоянии среда предоставляет агенту набор действий, выполнив которые он может повлиять на нее. В ответ на действия агента среда может менять состояния. За эту связь отвечает *функция перехода*. Среда может возвращать и сигнал вознаграждения. За эту связь отвечает *функция вознаграждения*. Совокупность этих двух функций называется *моделью* среды.

Цикл обучения с подкреплением



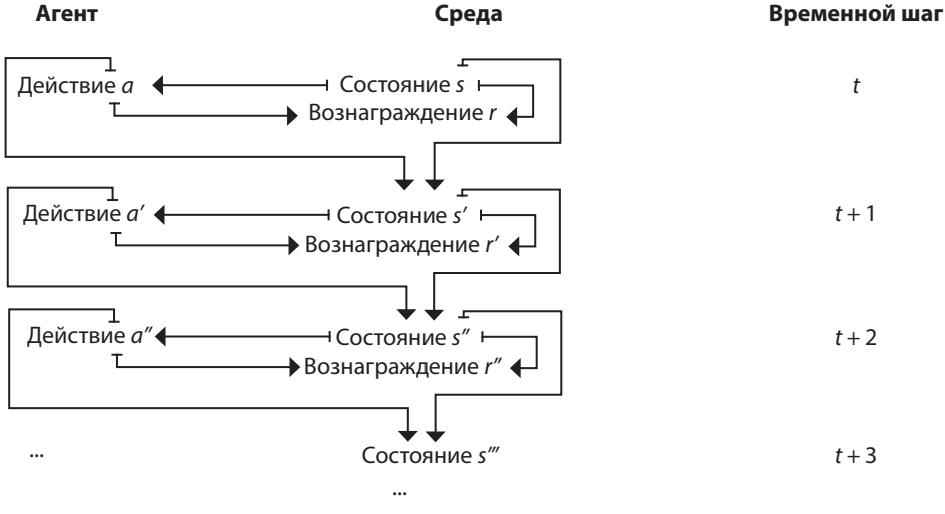
Обычно у среды есть четко определенная задача, которая формируется через функцию вознаграждения. Сигналы этой функции могут быть одновременно последовательными, оценочными и выборочными. Для достижения поставленной цели агент должен продемонстрировать интеллект или хотя бы когнитивные способности, связанные с ним: долгосрочное мышление, сбор информации и обобщение.

Работа агента проходит в три этапа: взаимодействие со средой, оценка ее поведения и улучшение ответов. Агент может быть предназначен для изучения связей между наблюдениями и действиями (такие связи называют *правилами* или *стратегиями*) или для изучения влияния *модели* среды. А с помощью *функций ценности* его можно научить оценивать и будущее вознаграждение.

Агенты глубокого обучения с подкреплением улучшают свое поведение методом проб и ошибок

Взаимодействие между агентом и средой продолжается несколько циклов — *временных шагов*. На каждом временном шаге агент наблюдает за средой, выполняет действие и получает новое наблюдение и награду. Совокупность состояния, действия, награды и нового состояния называется *опытом*. Любой опыт открывает возможность для обучения и улучшения производительности.

Кортежи опыта



Опыт:
t, (*s*, *a*, *r*′, *s*′)
t + 1, (*s*′, *a*′, *r*″, *s*″)
t + 2, (*s*″, *a*″, *r*‴, *s*‴)
...

...

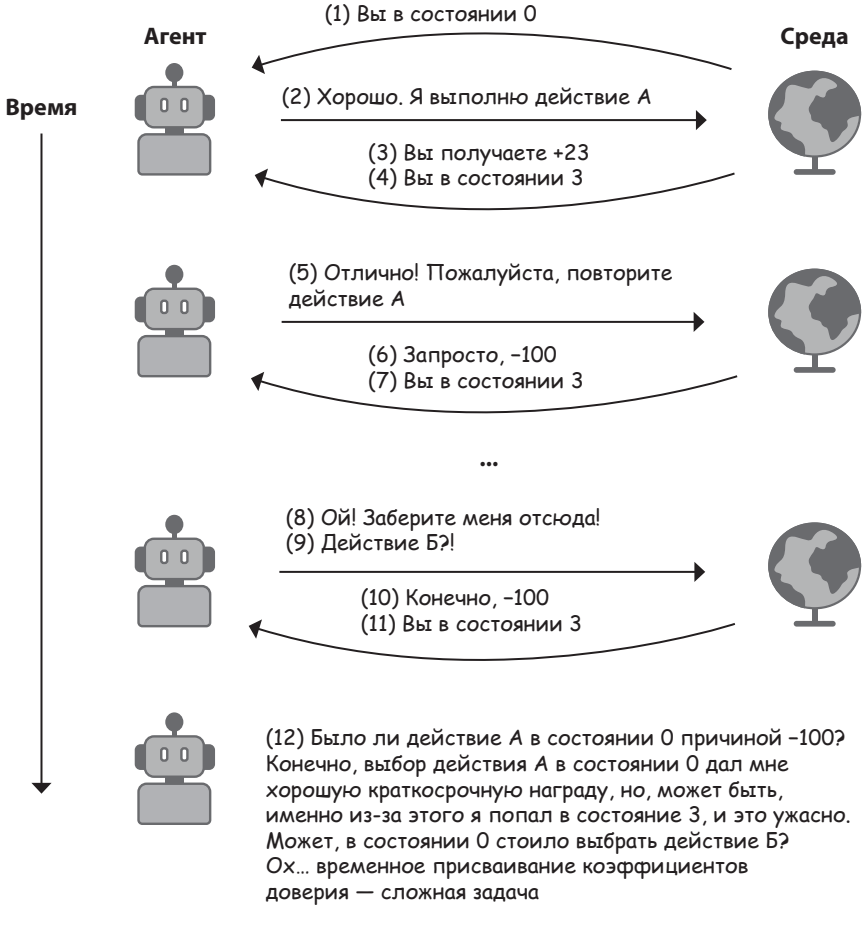
Задача, которую пытается выполнить агент, может иметь естественное завершение или не иметь его. Задачи с естественным завершением, такие как игры, называются *эпизодическими*, а без него, такие как обучение движению вперед, — *непрерывными*. Последовательность временных шагов от начала до завершения эпизодической задачи называется *эпизодом*. Чтобы научиться выполнять задачу, агенту может понадобиться несколько временных шагов и эпизодов. Агенты обучаются методом проб и ошибок: они пытаются что-то сделать, наблюдают результат, делают вывод, пробуют что-то другое и т. д.

Этот цикл мы подробно рассмотрим в главе 4 на примере среды с одношаговыми эпизодами. Начиная с главы 5, вы будете иметь дело со средами, которые требуют больше одного цикла взаимодействия в каждом эпизоде.

Агенты глубокого обучения с подкреплением учатся на последовательной обратной связи

У выполняемого агентом действия могут быть отложенные последствия. Вознаграждение может быть скудным и проявляться только через несколько временных шагов. Поэтому агент должен уметь учиться на последовательной обратной связи. Она лежит в основе так называемой *временной задачи присваивания коэффициентов доверия* — в определении того, какое состояние и/или действие привело к получению вознаграждения. Когда у задачи есть временная составляющая, а у действия — отложенные последствия, наградам сложно присвоить коэффициенты доверия.

Сложность временной задачи присваивания коэффициентов доверия

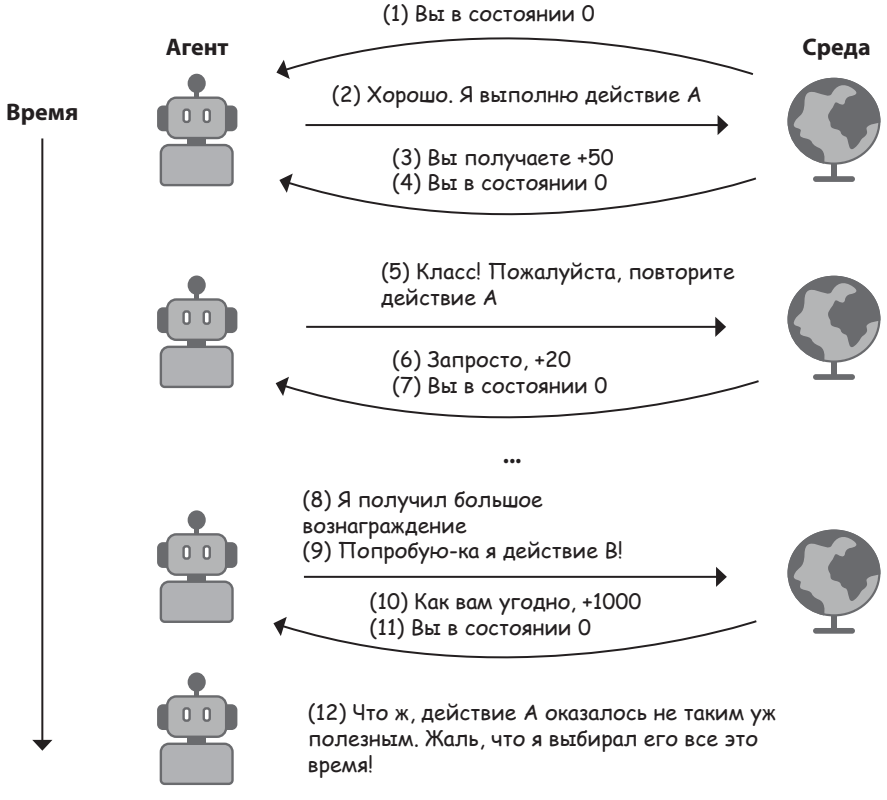


В главе 3 мы отдельно изучим все тонкости последовательной обратной связи. То есть ваши программы будут учиться на одновременно последовательной, контролируемой (в отличие от оценочной) и исчерпывающей (в отличие от выборочной) обратной связи.

Агенты глубокого обучения с подкреплением учатся на оценочной обратной связи

Агенту может быть недостаточно полученного вознаграждения — оно может не повлиять на процесс обучения. Награда может указывать на качество, а не на корректность: то есть она может не нести в себе информации о других потенциальных вознаграждениях. В связи с этим агент должен быть способен учиться на *оценочной обратной связи*. Такая обратная связь порождает потребность в исследовании. Агент должен уметь находить баланс между сбором новой информации и использованием уже имеющейся. Это называют *компромиссом между разведкой и эксплуатацией*.

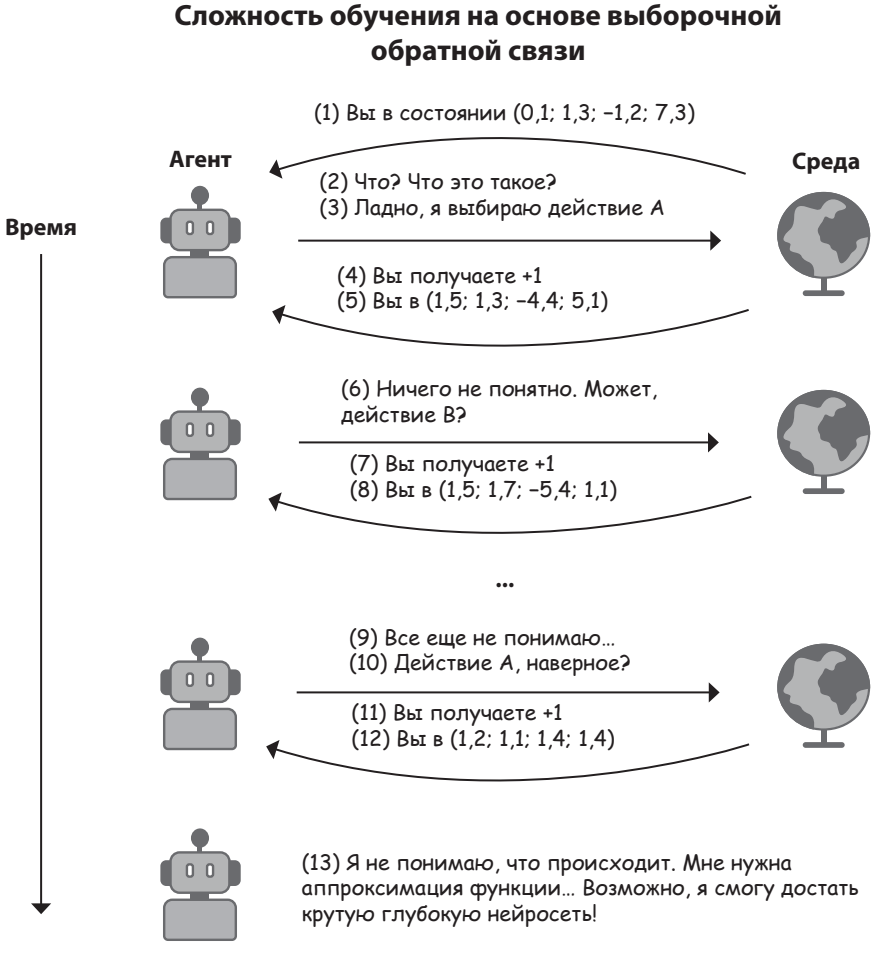
Сложность поиска компромисса между исследованием и эксплуатацией



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.

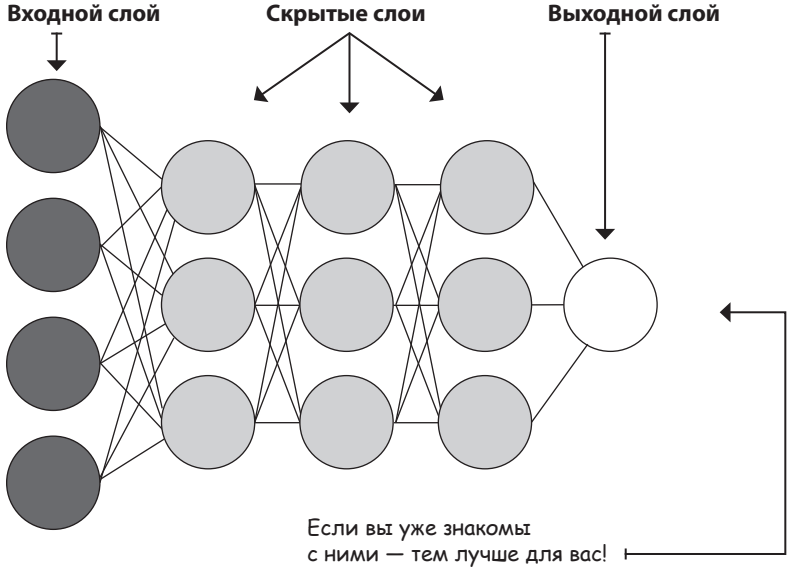


Агенты для аппроксимации правил называются *ориентированными на правила*, для аппроксимации функций ценности — *ценностно ориентированными*, для аппроксимации моделей — *модельно-ориентированными*, а агенты для аппроксимации и правил, и функций ценности называются *«актеры-критики»*. Агенты могут предназначаться 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