Потапов Алексей Сергеевич

Искусственный интеллект и универсальное мышление

Санкт-Петербург
2012

УДК 681.3
ББК 32.973
П64

Рецензенты: **В. В. Александров**, д-р техн. наук, профессор, заведующий лабораторией автоматизации научных исследований СПИИ РАН; **С. Н. Андрианов**, д-р физ.-мат. наук, профессор, заведующий кафедрой компьютерного моделирования и многопроцессорных систем СПбГУ

Потапов, А. С.
П64 Искусственный интеллект и универсальное мышление / А. С. Потапов. — СПб.: Политехника, 2012. — 711 с.: ил.
ISBN 978-5-7325-1008-9

Книга содержит доступное введение в обширную и сложную область искусственного интеллекта. Существенное внимание уделено основополагающим идеям, необходимым для глубокого понимания методов поиска в пространстве решений, представления знаний, машинного обучения и самоорганизации, составляющих основу искусственного интеллекта. В то же время книга представляет собой рассуждение о том, каких ключевых свойств не хватает интеллектуальным системам, чтобы стать по-настоящему разумными, для чего автор нередко обращается к истокам искусственного интеллекта в области психологии, лингвистики, нейрофизиологии, математики, философии.

Книга предназначена для широкого круга читателей, интересующихся вопросами мышления, но также может оказаться полезной и специалистам в области искусственного интеллекта.

УДК 681.3
ББК 32.973

ISBN 978-5-7325-1008-9

ОГЛАВЛЕНИЕ

Предисловие 6

Часть первая. **МЫШЛЕНИЕ КАК ПОИСК** 11
Големы и алхимики . . . . . –
От числа к алгоритму 16
Универсальная машина 23
Алгоритмическая неразрешимость 29
Проклятие размерности 38
Параллельность, непрерывность и квантовые компьютеры 43
Измерение интеллекта 52
Мыслят ли животные? 60
Мысль и действие 65
Рассуждения о методе 69
Эвристическое программирование 76
Направленный поиск 83
Успехи и неудачи эвристических программ 90
Общий решатель задач 100
Лабиринт алгоритмов 106

Часть вторая. **ПРЕДСТАВЛЕНИЕ ЗНАНИЙ** 115
Представление знаний как язык описания мира . . . . . –
Логика высказываний 120
Исчисление свойств и отношений 125
Знания, требующие рассуждений 132
Замкнутый мир и небытие 136
Неклассические логики 141
Знания в правилах 148
Нужна ли логика мышлению? 153
Власть над словом 157
N-граммы 165
Структура непосредственных составляющих 172
Формальные грамматики 175
Глубинные роли 184
Ассоциативные и семантические сети 189
Фреймы, объекты и программирование 196

Экспертные системы 207
От проблемы смысла к обучению 210

Часть третья. **МАШИННОЕ ОБУЧЕНИЕ** 220
Искусственные нейронные сети . . . . . –
Что значит учиться? . . . . . –
Загадка нервных клеток 226
Перцептрон 235
Модели ассоциативной памяти 243
От нейронов до зрения 250
Неоднозначность и адаптивный резонанс 260
Неклассические искусственные нейронные сети 274
Нейронные сети и машина Тьюринга 285
Распознавание и индукция 293
Распознавание образов и обучение . . . . . –
Соседи и эталоны 300
Математические методы распознавания 306
Распознавание и случайность 317
Обучение без учителя 328
Выбор признаков 335
Загадка индуктивного вывода 345
Что такое вероятность? 353
От случайности к информации 363
Алгоритмическая сложность и индивидуальная случайность 368
Универсальное пространство моделей 377
Принцип минимальной длины описания 387
Обучение и символы 401
Наука, красота и творчество . . . . . –
Выявление правил 412
Восстановление грамматик 421
Законы текста 429
Индукция и дедукция 442
Нечеткая логика 449
Теория свидетельств 461
Обучение понятиям 469
Обучение животных и модели рефлексов 483
Память 496
Воплощенный интеллект 507

Часть четвертая. **СТАНОВЛЕНИЕ ИНТЕЛЛЕКТА** 515
Интеллект и эволюция . . . . . –
Аниматы . . . . . –
Искусственные эмоции 523
Искусственная жизнь 538
Врожденность интеллекта 547
Эволюционные вычисления 566
Эволюция как поиск 579
Самоорганизация 595
Имитация отжига . . . . . –
Клеточные автоматы 600
Второе начало 608
Синергетика 617
Модели динамических систем 624
Детерминированный хаос 633
Самоорганизация в ИИ 646
Детерминизм, хаос и свобода воли 666
Проблема эмерджентности 672
Теория метасистемных переходов 681
Технологическая сингулярность 688

Заключение 701
Предметный указатель 708

ПРЕДИСЛОВИЕ

Загадка мышления — одна из самых захватывающих загадок в мире, ведь мышление — это то, что делает человека человеком. Оно не только отделяет людей как вид от животных (которые тоже в некоторой степени обладают мышлением), но и составляет основу нашей индивидуальности. Ведь что у человека нельзя отнять, не нарушив его личность? Утрата самой важной вещи может сказаться на личности лишь косвенно, хотя порой и весьма сильно. Даже утрата какой-то части тела, сколь бы печальной она ни была, не влияет напрямую на сохранность личности человека. Человек остается самим собой и при использовании донорской почки или искусственного сердца. И лишь о замене человеку мозга — органа, отвечающего за мышление, — говорить бессмысленно. Повреждения же мозга самым драматическим образом сказываются на личности.

Мы — это наше мышление. Не зря Декарт в своем великом сомнении оставил лишь один факт истинным: «Мыслю, следовательно, существую». Наличие мышления — единственный действительно бесспорный факт: даже чтобы сомневаться абсолютно во всем, нужно мыслить.

В то же время ребенок задает множество вопросов о внешнем мире, но почти никогда ничего не спрашивает о том, как и почему он думает. Да и любой школьник об устройстве мира может сказать гораздо больше, чем о своем мышлении, хотя именно последнее позволяет изучать все прочие загадки мироздания. И все же целью всего образования (по крайней мере, хорошего образования), в первую очередь, является развитие мышления и лишь во вторую очередь — приобретение конкретных знаний о мире. Почему же тогда ни в школе, ни в университете (за исключением отдельных специальностей) о мышлении практически ничего не говорится? А вдруг мы мыслим неправильно так же, как

неграмотный человек с ошибками использует язык? Ведь изучение родного языка (традиционно занимающее в школе многие годы) позволяет научиться не просто пользоваться им, но владеть. Также и изучение законов мышления могло бы поднять его использование на новый уровень. А в действительности оказывается, что даже многие ученые нередко хуже знают основной инструмент своего труда — ум, чем спортсмены — свое тело.

Причина, по которой с первых классов и на протяжении всего обучения в школе нет уроков по «основам мышления», заключается в том, что, хотя на протяжении веков механизмы мышления и интересовали ученых, но узнать о них удалось до обидного мало. Как заметил один видный психолог Э. Тулвинг: «Многие изобретения и открытия в других областях науки потрясли и озадачили бы Аристотеля, но самые яркие и неожиданные результаты психологических исследований... заставили бы его поднять брови только на мгновение». Конечно, данное утверждение сильно преувеличено, но определенная доля правды в нем есть.

Наиболее ярко недостаточность наших знаний о мышлении (наша «интеллектуальная безграмотность») проявляется при попытке его искусственного воспроизведения. Не зря говорится, что «машины не научатся думать до тех пор, пока не научатся думать люди». Полное понимание означало бы возможность создания мыслящей машины, искусственного интеллекта. Основные же классические определения мышления оказываются настолько поверхностными или неконкретными, что даже не позволяют разобраться, с чего нужно начать создание такой машины.

Подступиться к этой задаче удалось лишь с изобретением компьютеров. Многие проблемы мышления, о которых философы спорили веками, приобрели гораздо более ясный смысл (а некоторые и вообще исчезли) в контексте исследования искусственного интеллекта. И, наоборот, некоторые познавательные способности, которые выглядели элементарными, при попытке моделирования на компьютере нередко приводили к практически непреодолимым трудностям, вскрывающим всю глубину нашего непонимания данных процессов. К примеру, такой естественный и легкий для человека про-

цесс, как зрение, в действительности оказывается очень сложным и многогранным анализом информации. То, что наша зрительная система делает незаметно для нас, граничит с чудом. В то же время зрение традиционно полагалось всего лишь построением изображения. И лишь некоторые мыслители удивленно спрашивали: если глаз строит изображение, то кто же тогда дальше смотрит на это изображение, и что значит «смотреть»?

Конечно, это не означает, что сведения, накопленные в психологии, лингвистике и других когнитивных науках, связанных с изучением процессов познания, бесполезны. Напротив, именно там разработчики искусственного интеллекта черпают наиболее фундаментальные идеи. Просто эти идеи зачастую теряются в массе частных деталей, описывающих побочные продукты специфического человеческого мышления. Значительная часть психологии подобна той «информатике», которая посвящена изучению пакетов прикладных программ, имеющих лишь сиюминутное значение, вместо исследования того, что делает эти программы возможными.

Попытки воспроизведения мышления на базе компьютеров, длящиеся уже более полувека, тоже продвинулись как будто не слишком далеко. Сначала казалось, что до эры разумных машин остались считанные годы, ведь с помощью компьютеров очень быстро удалось сделать то, что раньше считалось невозможным, — передать машине умение играть в шахматы, решать математические задачи, распознавать некоторые классы образов и т. д. Однако время шло, а по-настоящему мыслящих машин не появлялось. Приутихшие высказывания о невозможности создания мыслящих машин зазвучали с новой силой. И даже термин «искусственный интеллект» (ИИ) стал использоваться для обозначения не гипотетической разумной машины, а области научных исследований и технических разработок, посвященных проблемам построения «интеллектуальных» систем, заменяющих человека при решении конкретных задач или моделирующих внешние проявления отдельных мыслительных операций.

Сейчас искусственный интеллект является в большей степени метафорой в том смысле, что, как отмечает профессор

В. В. Александров, в применении к разным областям знаний (экономике, системам управления, компьютерному зрению и т. д.) интеллектуальные методы опираются на разный математический аппарат. Как результат, вместо одной области исследования появляется множество мало связанных направлений, число которых увеличивается с расширением сфер применения автоматических методов. Во многом с этим можно согласиться. Однако все же можно надеяться найти во всех этих методах некую фундаментальную общность или попытаться построить систему их автоматического синтеза.

К сожалению, в литературе по ИИ фундаментальные проблемы мышления тоже тонут в море технических деталей. Многие учебники по ИИ приобретают оттенок схоластичности: в них пересказываются стандартные утверждения, читаемые как заклинания; они выглядят очень важными, но их связь с естественным интеллектом теряется, смысл ускользает, и остается лишь повторять их для следующих поколений в надежде, что у кого-то то или иное заклинание сработает. Число работ в этой области не уменьшается, а даже наоборот, возрастает, но уже почти никто не заявляет о цели своих исследований как о создании мыслящей машины. В основном ставятся более скромные, более наукообразные или более финансируемые цели, достижение которых прогнозируемо.

Некоторые исследователи не отказываются от глобальной цели создания мыслящей машины, которую стали называть универсальным, или *сильным ИИ*, противопоставляя его *слабому ИИ*, способному решать отдельные интеллектуальные задачи, но не обладающему той универсальностью, которой обладает человеческий разум. Однако эти исследователи обычно пытаются найти нетрадиционные пути достижения данной цели, утверждая, что вся область ИИ находится в тупике, и во многом отрицая ее предыдущие достижения. Компьютерам, действующим по строгим программам, часто отказывают в свободе воли, способности к творчеству, самосознанию и многим другим особенностям человеческого мышления. Говорят, что компьютеры не могут сделать ничего сверх того, что в них заложено человеком. В этой связи выглядит парадоксальным то, что именно компьюте-

ры оказались столь успешным (и, по сути, единственным) средством моделирования мышления.

Цель настоящей книги — разрешить этот парадокс и показать, что никакого тупика в классическом (алгоритмическом) подходе к ИИ, в действительности, нет. Хотя начальные ожидания в этой области оказались сильно завышенными, в ней идет постоянное развитие, основные результаты которого как кусочки мозаики, постепенно образуют целостную картину феномена мышления. В этой картине остается все еще много пробелов, которые, однако, можно увидеть, только соединив уже имеющиеся части мозаики, что мы и попытаемся сделать, не слишком углубляясь в технические детали, но и не теряя сути (из-за чего, правда, придется немного пожертвовать популярностью изложения).

В связи с этим книга имеет во многом обзорный характер, и, хотя ее нельзя рассматривать как учебник, она может использоваться в качестве введения в проблематику ИИ для студентов и молодых ученых. Взгляд на человеческое мышление с перспективы ИИ, затрагиваемый в книге, может оказаться интересен для широкого круга читателей. В то же время обсуждение научно-методологических вопросов данной области может представлять интерес и для специалистов, проводящих исследования в сфере когнитивных и компьютерных наук.

Автор выражает благодарность С. А. Родионову, Н. И. Потаповой, А. И. Свитенкову и А. Г. Мясникову за предварительное знакомство с книгой и высказанные замечания. Автор также признателен участникам научного семинара «Искусственный интеллект: от методологии к инновациям», проводимого на базе кафедры компьютерной фотоники и видеоинформатики НИУ ИТМО, интернет-форума ailab.ru (ранее aicommunity) за многолетнее плодотворное обсуждение проблем искусственного интеллекта.

Уважаемый читатель!

Свои отзывы и пожелания присылайте автору по e-mail: pas.aicv@gmail.com

Часть первая

МЫШЛЕНИЕ КАК ПОИСК

ГОЛЕМЫ И АЛХИМИКИ

Термин «искусственный интеллект» (ИИ) уже давно появляется чаще в научной литературе, чем в фантастике. Собираются многотысячные симпозиумы, выпускаются книги, журналы и даже учебники, озаглавленные «Искусственный интеллект». Можно ли ожидать в них увидеть инструкцию по созданию ИИ? Непосвященный человек, видимо, весьма удивится, ознакомившись с данными материалами и не найдя почти никакой связи с бытовым понятием ИИ, почерпнутым из художественных книг и фильмов. Сейчас ИИ — это, скорее, наименование большого направления научных исследований, связанных с решением частных задач автоматизации и с моделированием отдельных элементов человеческого мышления. Как ни странно, почти нет специалистов в области ИИ, которые бы сказали, что занимаются созданием настоящего (сильного или универсального) искусственного интеллекта. Немалое их число может даже отрицать саму возможность его существования, сравнивая искусственный интеллект с философским камнем в алхимии.

На заре возникновения области ИИ настроения ученых были совершенно иные. Многими энтузиастами полагалось, что создание настоящего искусственного интеллекта, не уступающего человеческому, — дело нескольких десятилетий. Эти ожидания не оправдались, что нередко рассматривается как неудача всего направления исследований. «Профессионалы» теперь не питают наивных иллюзий о реалистичности собственноручного создания сильного ИИ и

руководствуются в своей работе прагматическим подходом, занимаясь разработкой самых разнообразных прикладных «интеллектуальных технологий» (слабого ИИ), образующих столь пеструю мозаику, что в ней сложно увидеть какую-то общую картину. И, тем не менее, имеется глубокая связь всех исследований в области ИИ, как бы представляющих проекции одного интеллекта на разные плоскости. Стоит попытаться восстановить имеющуюся сейчас картину, начиная с самого начала.

Обычно момент начала разработок в области искусственного интеллекта относят к 1950-м годам — немногим после создания первых компьютеров. Считается, что само понятие ИИ было закреплено в 1956 году в названии семинара, проходившего в Дартмутском колледже, а в 1960-х годах приобрело широкое распространение. В 1969 году прошла 1-я Международная объединенная конференция по искусственному интеллекту. Но неужели до этого люди не задумывались о возможности воссоздания разума?

К примеру, не секрет, что термин «робот» был впервые использован чешским писателем Карелом Чапеком в пьесе «Р.У.Р.» 1921 года, а мифы об искусственно созданных человекоподобных существах уходят в гораздо более глубокое прошлое. Так, в средневековой Европе алхимики пытались создать гомункулусов — искусственных существ, подобных человеку. Еще раньше возникли мифы об оживлении големов (существ из неживой материи) с помощью каббалистической магии. Легенды об искусственных существах есть у многих древних народов.

Самостоятельно вдохнуть жизнь в мертвую материю!.. Это было вполне естественной мечтой, ведь именно так люди могли сравниться со своими богами. И по этой же причине подобные желания часто рассматривались как еретические, а искусственные существа считались уделом черной магии. И до сих пор можно услышать религиозные возражения против возможности создания искусственного интеллекта.

Однако в пользу воспроизведения, по крайней мере, некоторых способностей живых организмов говорили чисто практические соображения. По сути, вся история техники — это история замены или дополнения живого искусственным:

колесо вместо ног, рычаг вместо мускулов. Но подобные изобретения — лишь инструменты, которыми управляет человек. Сами по себе они не проявляют свойств живых организмов, не обладают собственным поведением.

Неужели не было попыток разработки «самостоятельных» механизмов? Ведь в сказках всегда встречаются так желанные людьми скатерть-самобранка или ковер-самолет и многие еще более самостоятельные вещи. Но сказку никак не получалось превратить в быль. На практике созданию таких предметов мешало то, что мышление традиционно считалось чем-то нематериальным, а значит, неподвластным для воспроизведения в обычной технике. Даже орган мышления долгое время не получалось точно установить: им считались то сердце, то печень. Тем не менее, не только вымышленные, но и вполне реальные искусственные существа имеют давнюю историю. В разных странах и в разные времена были созданы многочисленные механические птицы, музыканты, играющие на различных инструментах, танцовщицы и прочие разнообразные механические игрушки. Попытки автоматизировать ремесленный труд тоже проводились задолго до появления компьютеров. Разного рода станки особенно широко стали распространяться в XVIII веке. Все эти механизмы совершали какие-то самостоятельные действия, чем заметно отличались от обычных инструментов, оживающих лишь в руках человека.

Однако живое от обычной техники всегда отличала не только собственная активность, но также сложность поведения и способность реагировать на внешние воздействия. Представьте себе автомат, выполняющий жестко заданные действия вне зависимости от их целесообразности, к примеру, штамповочный автомат на пустом конвейере. Вряд ли он будет восприниматься живым. Напротив, почти живым будет казаться, скажем, робот-игрушка, не выполняющий никаких утилитарных функций, но реагирующий на ситуацию так же, как и домашнее животное. Способность к такой реакции, характерная для любого живого организма, означает наличие мышления или каких-то его зачатков.

Что значит думать? Думаете ли вы все время? Чем занят ваш мозг? Давайте посмотрим, к каким ситуациям применим

глагол «думать». Обычно говорят, что человек думает над какой-то задачей или проблемой: ученик думает над вопросом учителя; философ думает о проблеме бытия; шахматист думает над игровой задачей. В качестве предмета мышления может выступать какая-то загадка, вопрос, школьная задачка, ситуация в некоторой интеллектуальной игре, жизненная проблема и т. д. Во многом мышление — это процесс решения проблем. Когда кто-то слишком быстро отгадывает какую-то загадку, ему могут возмущенно сказать: «ты знал!», — поэтому если для какой-то задачи ответ заранее известен, мышление особо и не нужно. Другими словами мышление «включается» в ответ на задачу, проблемную ситуацию, для которой нет готового решения.

Чтобы техника воспринималась хоть немного «живой», она должна уметь правильно разрешать новую для себя ситуацию, т. е. обладать хоть какими-то зачатками интеллекта. Сколько бы изобретатели ни старались, им никак не удавалось достичь сходства в поведении своих детищ хотя бы с животными, не то что с человеком. Некоторые шли на прямой обман. Наиболее известным случаем является шахматный аппарат Кемпелена, внутри которого прятался живой человек. Этого «искусственного» игрока до разоблачения успели продемонстрировать многим правителям, в числе которых был и Наполеон. Справедливости ради нужно отметить, что Вольфганг фон Кемпелен был изобретателем и ряда реально действующих машин, в частности, «говорящей» машины, имитировавшей голос ребенка с помощью системы паровых клапанов.

Предпринимались попытки создания машин, которые бы помогали человеку и в умственной деятельности. Проще всего автоматизировались арифметические операции, которые к тому же были весьма востребованными. Первые арифметические машины, видимо, появились еще до нашей эры. Однако наиболее известными являются машины Блеза Паскаля, построенные им на основе часового механизма в XVII веке.

В те времена умением считать обладал не каждый человек, и, конечно, это умение тогда еще не усматривалось у животных. Не удивительно, что арифметические вычисления признавались примером сложной умственной деятельности (это сейчас мы их считаем не особо интеллектуальным занятием; в противном случае мы бы вынуждены были признать, что компьютеры в чем-то умнее нас). Возможность автоматизации элементов умственной деятельности, не доступных животным, позволила Паскалю высказывать весьма смелые для XVII века мысли о возможности механического воспроизведения мышления в целом. Ведь идея небожественности мышления могла показаться гораздо более кощунственной, чем, например, идея гелеоцентризма. Еще раньше Рене Декарт рассуждал о человеке как о машине (при этом он, правда, не отрицал существования отдельной «мыслящей субстанции»).

Помимо арифметических машин предлагались устройства для автоматизации и других форм умственной деятельности, например машины для поиска книг в библиотеках или машины для сравнения идей («идеоскопы» С. Н. Корсакова).

Казалось бы, все эти вполне удачные машины являются прообразами современных бытовых и промышленных роботов, и как раз их развитие должно было привести к возникновению области искусственного интеллекта. Однако увлечение «механической жизнью» закончилось в XIX веке, наткнувшись на непреодолимые препятствия.

Причина заключалась не столько в малой пригодности механики для реализации сложных систем управления, сколько в том, что изобретатели имитировали лишь особенности внешнего поведения без понимания обеспечивающих его внутренних процессов. Каждый механизм был уникальным произведением технического искусства (слово «техника» в древности как раз и означало «искусство» или «мастерство»). Даже более сложные человекоподобные устройства, сконструированные в 1920-х годах на новой элементной базе, не преодолели этого ограничения. И лишь с изобретением компьютеров стало возможным появление науки об искусственном интеллекте.

Но разве основой современных компьютеров не являются арифметические машины, которые существовали уже многие века? Однако подумайте, что для вас компьютер? Что является его основной отличительной особенностью? Несмотря на то, что многие люди не понаслышке знакомы

с потенциально почти безграничными возможностями использования компьютеров, до сих пор бытует мнение, будто компьютер — это просто большой калькулятор, основная задача которого — считать.

Это серьезное заблуждение. Будет не слишком большим преувеличением сказать, что изначально компьютеры разрабатывались с единственной целью — моделировать мышление, а не производить вычисления. И хотя изобретение разнообразных механических устройств сыграло в этом определенную роль, появлению компьютеров мы обязаны, в первую очередь, весьма абстрактным теоретическим работам математиков.

ОТ ЧИСЛА К АЛГОРИТМУ

Прежде чем обсуждать, в чем же заключается фундаментальный вклад математики в возникновение компьютеров, попробуем подумать над вопросом: что такое математика? Что в ней является предметом исследования? Первое, что приходит на ум, — это числа, или количественные отношения. Но что же это такое — число?

Понятие числа удивляет своей универсальностью. Два плюс два равно четырем вне зависимости от того, складываем ли мы яблоки или секунды, выполняем ли мы сложение на Земле или на Марсе. Сложно представить себе мир, в котором результат сложения менялся бы каждый раз. И даже фантасты, несмотря на все свое воображение, не рискуют описывать подобные миры! Такое постоянство результатов сложения связано, в первую очередь, с тем, что сложение — это умственная операция объединения объектов, не подразумевающая какого-то конкретного физического взаимодействия между ними. Ведь «физическое» сложение далеко не всегда равносильно математическому, и если сложить вместе две половинки критической массы радиоактивного вещества, в ответе получится не просто единица...

И все же математическое сложение имеет подозрительно много применений в реальном мире. Так, первобытный человек мог загибать и разгибать пальцы, считая входящих в пещеру и выходящих из нее соплеменников, и знать число людей в пещере. Вдумайтесь: он мог предсказывать нечто, не видя это собственными глазами. Такие способности были настоящим волшебством, которому сейчас учат с малых лет!

Как можно догадаться заменить отсутствующие объекты загнутыми пальцами или отложенными камушками, причем поняв, что такую операцию можно проводить независимо от типа объектов? Нетривиальность подобных открытий часто недооценивается. На самом деле, на формирование современного понятия числа ушло значительное время. Во многих языках сохранились следы промежуточных этапов этого процесса, на которых, в частности для обозначения количеств объектов разных типов, использовались разные числительные. Например, в японском языке (как и в ряде других языков) остались счётные суффиксы, различающиеся для предметов разных форм и типов (при счете вытянутых предметов к числу добавляется суффикс «хон», плоских предметов — «бон», больших предметов и единиц техники — «дай» и т. д.).

С «изобретением» чисел как самостоятельных объектов, которым могут быть приписаны собственные свойства, начался период элементарной математики. В изучении свойств чисел особенно преуспела пифагорейская школа. Помимо количественных отношений исследовались также и пространственные отношения между объектами. Как и в случае чисел, свойства геометрических фигур не зависят от того, на каком материале и с помощью какого инструмента они нанесены. Абсолютность и неизменность математических отношений даже заставили Платона и многих последующих мыслителей увериться в том, что этот идеальный мир первичен (и в нем содержатся не только числа, но и идеальные образы моральных, этических понятий и категорий мышления), а материальная реальность — лишь искаженное, неточное его воспроизведение.

Математика пошла дальше по пути абстрагирования не только от реальных объектов, заменяя операции с ними действиями над числами, но и от самих чисел, перейдя к операциям с переменными величинами. Например, мы

можем написать равенство $A + B = B + A$, выполняющееся независимо от конкретных значений A и B. Это свойство самой операции сложения: для деления это равенство уже не будет выполняться. Более сложным примером является вывод решения квадратного уравнения. Работа с переменными величинами, поиск для них общих законов — это удел высшей математики, основное развитие которой приходится на XVII–XVIII века. Числа постепенно исчезают из чистой математики и остаются лишь символы.

Это достижение является крайне важным. Попробуйте, к примеру, не прибегая к символьным обозначениям, решить задачу: «Какова сторона квадрата, площадь которого на двенадцать больше его периметра?» Решение этой задачи в словесной форме заняло бы много места, и его достоверность могла бы быть под сомнением. Кроме того, символьная запись позволяет решать подобные задачи в общем виде, заменяя конкретные численные значения в условиях задачи новыми переменными. Получив общее решение, нам достаточно будет в него просто подставлять конкретные числа вместо того, чтобы каждый раз заново решать одну и ту же по форме задачу.

Однако современная математика и на этом не останавливается. В ней происходит полное абстрагирование от «смысла» как математических объектов, так и операций над ними. Нет больше чисел и треугольников, сложения и умножения, есть лишь множества объектов и абстрактные операции над ними, введенные через явное указание всех свойств операций (т. е. через системы аксиом).

В отличие от аксиом в геометрии Евклида, здесь природа объектов не уточняется, а выбор аксиом произволен. Можно подобрать аксиомы так, чтобы свойства заданного множества объектов совпадали с привычными для нас свойствами чисел. Но числа — лишь один из бесконечного разнообразия типов объектов, которые можно задать аксиоматически. Конечно, остается вопрос, какие системы аксиом задают интересные множества — те множества, которые выбираются математиками для детального анализа.

Одна из возможностей — задавать наименьшее число аксиом, которые все еще позволяют получить нетривиальные

следствия. К примеру, можно ввести произвольное множество объектов G и задать на этом множестве некоторую операцию #, для которой выполняются следующие аксиомы:

1) для любых A, B, C из G верно $(A \# B) \# C = A \# (B \# C)$;

2) существует такое E из G, что для любого A из G верно $E \# A = A \# E = A$;

3) для любого A из G существует Z из G такое, что $A \# Z = Z \# A = E$.

Следует оговориться, что аксиомы 1–3 могут быть записаны вовсе без слов, с использованием нескольких дополнительных символов (символа принадлежности множеству, кванторов существования и всеобщности).

Такое множество будет называться группой (E называется единичным элементом, а Z — обратным к A элементом). Из аксиом можно получить разнообразные следствия. К примеру, можно доказать, что обратным элементом к единичному является сам единичный элемент или что если $A = B$, то $A \# C = B \# C$ для любого C.

Большое число следствий из данного набора аксиом, описывающих свойства групп, можно перенести на самые разные объекты и операции. Как результат, не нужно будет одни и те же по форме выводы повторять для каждого нового типа объектов. Так же, как подстановка конкретных чисел в формулу дискриминанта позволяет сразу получить решение квадратного уравнения, подстановка объектов конкретного типа в одну из подходящих систем аксиом позволяет сразу узнать для них большое число свойств. В частности, группой является множество целых чисел с операцией сложения. Множество преобразований подобия на плоскости с операцией суперпозиции тоже образуют группу. Множества преобразований, например сдвига или поворота системы как целого в пространстве, относительно которых физические законы симметричны (инвариантны), также математически являются группами. На все эти множества можно перенести свойства групп, выведенных всего из нескольких аксиом. Дальнейшее расширение системы аксиом позволяет уточнить типы объектов, для которых выводится большое число более конкретных свойств.

Можно согласиться, что аксиоматический подход выглядит весьма полезным, но при чем же здесь компьютеры и

искусственный интеллект?.. Дело в том, что в математике в центре внимания всегда была проблема истинности. С помощью рассуждений мы проводим доказательство истинности некоторой теоремы. Но где гарантия, что проводимые людьми рассуждения корректны? И, кроме того, как человек придумывает доказательства? «Человеческий фактор» всегда мешал математике достигнуть идеальности. В связи с этим и возникал вопрос: как математически описать само мышление человека?

Не случайно логика как наука о правильном мышлении, возникшая еще несколько тысячелетий назад вслед за арифметикой и геометрией, сейчас является разделом математики. Логика во многом стала основой аксиоматического подхода, поскольку предоставила методы получения истинных следствий из истинных фактов и общих правил (аксиом). Иными словами, аксиоматический подход возник в результате попытки избавиться от «человеческого фактора» в математике, строго описать процесс решения человеком математических проблем, среди которых многие все еще оставались открытыми.

Список наиболее сложных нерешенных проблем был составлен выдающимся математиком Давидом Гильбертом в 1900 году. Некоторые из этих проблем могут показаться на первый взгляд простыми. Например, десятая проблема Гильберта заключалась в разработке метода нахождения решений (или доказательства их отсутствия) для систем диофантовых уравнений. Диофантово уравнение является алгебраическим уравнением, у которого как все коэффициенты, так и все значения неизвестных являются целочисленными. Например, следующая система диофантовых уравнений: $x \cdot y \cdot z = 105$ и $x - y = 4$ имеет решение $x = 7$, $y = 3$, $z = 5$. Но как разработать метод, «механическое» применение которого будет позволять решать диофантовы уравнения?

В 1928 году была сформулирована «программа Гильберта» по обоснованию математики. Эта программа восходит к рассуждениям Декарта о методах решения математических проблем. Чуть позже (еще в XVII веке) и Лейбниц после создания арифметической машины, аналогичной машине Паскаля, размышлял о возможности построения машины,

которая бы, манипулируя символами, «вычисляла» истинность математических утверждений. Необходимость обоснования математики стала особенно важной после того, как в конце XIX века в основах математики (в частности, в теории множеств Кантора) обнаружились противоречия. А как известно из логики, из противоречивых посылок можно вывести произвольные следствия, т. е. многие из полученных ранее математических результатов вполне могли оказаться ошибочными! Аксиоматический подход (наряду с интуиционализмом) стал одним из подходов к устранению подобной опасности. Программа Гильберта по доказательству непротиворечивости классической математики подразумевала ее полную формализацию через построение системы аксиом и правил вывода, с помощью которых можно было бы получить все основные математические теоремы. При этом доказательства должны были состоять из таких элементарных шагов, корректность которых проверялась непосредственно. Это привело к формализации понятия алгоритма.

Само слово «алгоритм» возникло гораздо раньше. Оно происходит от Algorithmi — латинского написания имени арабского математика аль-Хорезми. В первой половине IX века появился его трактат с описанием разработанной в Индии десятичной системы счисления. Благодаря переводу этого трактата в XII веке на латинский язык средневековая Европа познакомилась с позиционной системой счисления, поэтому долгое время алгоритмом называлось искусство счета в этой системе. Интересно, что от арабского названия этой книги также происходит слово «алгебра».

Вычисления в позиционной системе счисления описывались гораздо более простыми и единообразными правилами, которые могли быть легко использованы (попробуйте, к примеру, произвести элементарные арифметические операции в непозиционной римской записи: XIX + VI или XII – IV, — без их перевода в десятичную систему). Именно благодаря этому свойству вычисления могли быть воплощены в машинах Паскаля, являвшихся аппаратной реализацией соответствующих алгоритмов.

Начиная с XII века, значение слова «алгоритм» постепенно расширялось, пока не стало означать произвольную четко

определенную последовательность действий над некоторыми объектами, позволяющую решать задачи некоторого типа (вовсе не обязательно связанные с вычислениями). Ретроспективно алгоритмами оказались и разнообразные методы построения геометрических объектов с помощью циркуля и линейки, и метод Евклида определения наибольшего общего делителя двух чисел. Последний можно описать так: «Для определения наибольшего общего делителя двух положительных целых чисел следует последовательно вычитать из большего числа меньшее, пока они не сравняются. Полученное число и есть ответ».

В виде алгоритмов представлялись методы решения самых разнообразных математических задач, но сами алгоритмы оставались нематематическими объектами, поскольку составлялись преимущественно из фраз естественного языка. По сути, десятая проблема Гильберта состоит в поиске алгоритма решения системы диофантовых уравнений. Неизбежно возникал вопрос: как строго определить понятие алгоритма? Важнейшим достижением математики XX века стала формализация понятия алгоритма в рамках аксиоматического подхода.

Современная математика в значительной степени становится наукой о формальных системах. Математические объекты и операции над ними полностью заменяются символами. Цепочки символов описывают математические выражения. Допустимые преобразования цепочек также описываются символьно, в рамках выбранного набора аксиом. В результате решение математической задачи представляется как последовательность преобразований цепочки символов от исходных данных к решению. Такая последовательность задается алгоритмом, который также является математическим символьным объектом.

Существование формального описания некоторого алгоритма крайне важно, поскольку означает возможность «механического» повторения этого алгоритма так же, как алгоритмы сложения и умножения чисел воспроизводились машиной Паскаля.

Любые ли алгоритмы имеют формальные описания? Для любой ли задачи существует алгоритм ее решения? Какой

архитектурой должно обладать гипотетическое устройство, выполняющее некоторый алгоритм?

УНИВЕРСАЛЬНАЯ МАШИНА

Одно из первых строгих определений понятия алгоритма предложил Курт Гёдель. Опираясь на свое определение, в 1931 году он доказал существование алгоритмически неразрешимых математических проблем, т. е. таких проблем, для которых отсутствуют алгоритмы их решения. Для математиков это было настоящим потрясением, поскольку говорило о невыполнимости программы Гильберта и разрушало мечту о достижении абсолютной строгости в математике.

Оставалась еще надежда, что существование алгоритмически неразрешимых проблем связано с тем, что Гёдель предложил неудачное определение, описывающее слишком узкий класс алгоритмов. Были предприняты попытки дать наиболее общую формализацию понятия алгоритма, в рамках которой можно было бы найти алгоритм решения любой математической проблемы.

Дальнейшее уточнение понятия алгоритма предложил в 1936 году Алан Тьюринг (похожее уточнение практически одновременно с Тьюрингом дал Эмиль Пост). Их формализмы представляли собой некие гипотетические устройства (машины), реализующие автоматический процесс обработки символьной информации. Были предложены и другие формальные модели алгоритма, например лямбда-исчисление Алонзо Чёрча (который даже немного опередил Тьюринга и Поста), нормальные алгорифмы А. А. Маркова, рекурсивные функции Эрбрана и Гёделя, нормальные системы Э. Поста. Все эти формализмы оказались эквивалентными в том смысле, что математические проблемы, обладающие решениями в рамках одного формализма, также обладают решениями и в рамках остальных формализмов. Неудачи в попытке дальнейшего расширения понятия алгоритма привели к принятию тезиса Чёрча—Тьюринга, который гласит, что понятие *машины Тьюринга* или любое из эквивалент-

ных существующих определений полностью описывает интуитивное понятие алгоритма и не может быть расширено. В терминах вычислимости говорится, что любая функция, вычислимая в каком-либо естественном смысле, вычислима с помощью машины Тьюринга (т. е. существует программа, которая, получив на вход значение аргумента, напечатает на выходе значение функции).

После формализации понятия алгоритма произошла его столь широкая экспансия во все сферы человеческой деятельности, что это понятие стало одной из основополагающих категорий. В том числе и исследования человеческого мышления сейчас трудно представить без использования этого понятия. В результате его границы стали расширяться, и сейчас порой говорят, будто математическое определение алгоритма не соответствует его интуитивному пониманию, хотя даже само слово «алгоритм» возникло именно в математике. Кроме того, еще полвека назад в энциклопедиях это понятие если и присутствовало, то только в качестве сугубо специального термина.

Идея алгоритмов связана, в первую очередь, с формализацией процесса решения математических проблем, а современная математика — это наука о символьных системах. В связи с этим машина Тьюринга представляет собой модель устройства, имитирующего элементарные операции человека над строками символов в процессе математических рассуждений.

В общем виде машина Тьюринга — это абстрактное устройство, которое в каждый момент времени находится в одном из конечного числа внутренних состояний, а также обладает бесконечной лентой, реализующей внешнюю память. В каждой клетке ленты, на которые она разделена, может быть записана любая из букв некоторого (фиксированного для данной машины Тьюринга) алфавита. Также машина Тьюринга снабжена головкой, движущейся по ленте и способной читать и записывать символы в клетки ленты. Функционирование машины описывается таблицей переходов, состоящей из совокупности команд. Каждая команда содержит условие применения, устанавливающее, при каком текущем внутреннем состоянии и каком символе в

текущей ячейке ленты данная команда может применяться. Если условие выполняется, то в соответствии с выбранной командой машина может одновременно осуществить переход в новое внутреннее состояние, записать в текущую ячейку ленты новый символ и передвинуть головку на новую ячейку. Таблица команд называется программой для машины Тьюринга.

При начале работы машины Тьюринга ее внутреннее состояние сбрасывается в выделенное (начальное) состояние, а головка помещается в начало ленты. Исходное содержимое ленты, однако, задается извне. Машина Тьюринга, руководствуясь программой, перерабатывает содержимое ленты. Если для текущей конфигурации машины (конфигурация определяется текущим внутренним состоянием, содержимым ленты и положением головки) нет подходящей команды, машина останавливается. Существуют также некоторые внутренние состояния, которые считаются конечными, — попадая в них, машина также останавливается.

Таким образом, машина Тьюринга получает на вход некоторую строку, перерабатывает ее и останавливается, давая на выходе новую строку. Исходно подразумевалось, что содержимое ленты — это математические выражения. Начальное состояние ленты — это формулировка условия задачи некоторого класса. Программа задает алгоритм решения задач данного класса. Например, не слишком сложно предложить машину Тьюринга для вычисления результатов арифметических операций в десятичной системе счисления (входное состояние будет соответствовать выражению, а выходное — его значению). Другая машина Тьюринга может быть предложена для поиска общего делителя двух чисел или для установления истинности логических выражений, или для решения какой-либо иной задачи.

Машины Тьюринга с разными программами — это разные воображаемые устройства. Если бы мы захотели их воплощать аппаратно, то под каждую машину пришлось бы создавать собственный механизм наподобие машины Паскаля. Именно поэтому никакое развитие физики с появлением новой элементной базы (от шестеренок до транзисторов) само по себе не привело бы к возникновению компьютеров: просто

появлялись бы все более эффективные специализированные машины типа калькуляторов.

Современный человек, хоть немного знакомый с программированием, да и просто с компьютерами, может воскликнуть: но почему математики не догадались саму программу тоже кодировать строкой символов и записывать на входную ленту?! К счастью, они догадались, иначе компьютеров не было бы до сих пор, но догадаться было заметно труднее, чем это может показаться сейчас, когда такая возможность широко известна и ее надо не открывать, а лишь обратить на нее внимание.

Да, мы можем представить программу как цепочку символов. Ведь именно для достижения этой цели и осуществлялась формализация понятия алгоритма. Но существует ли такая машина, которая принимала бы на вход строку, составленную из программы (описания конкретной машины) и данных, и могла бы получить на выходе ту же строку, какую бы получила описанная в начале ленты машина? Существование такой машины вовсе не самоочевидно, ведь, по сути, она должна уметь превращаться в абсолютно любую другую машину, будь то машина Паскаля или идеоскоп Корсакова. К счастью, Тьюрингу удалось провести доказательство существования такой *универсальной машины* и найти конкретные ее реализации. Исходно это была абсолютно абстрактная проблема, в решении которой было заинтересовано лишь небольшое число математиков. Однако сейчас любой процессор компьютера с внутренними регистрами и внешней памятью, в которую могут записываться любые программы и данные, представляет собой физическое воплощение машины, эквивалентной универсальной машине Тьюринга. Таким образом, компьютер не выполняет вычисления, а воспроизводит действие произвольной мыслимой машины, описание которой подается ему на вход в виде символьной строки.

Оказывается, число шагов, необходимых универсальной машине для имитации любого шага конкретной машины, является функцией, полиномиально зависящей от длины программы, т. е. универсальная машина не принципиально проигрывает в быстродействии специализированным машинам. То, что эта зависимость не является экспоненциальной,

важно при рассмотрении вопросов, касающихся сложности вычисления (о чем будет сказана несколько позднее). Кроме того, универсальные машины, на которых можно выполнить любой алгоритм, оказываются весьма простыми и могут иметь малое число собственных внутренних состояний.

То препятствие, на которое наткнулись средства автоматизации в XIX веке, — это отсутствие теории, которая бы позволяла исследовать алгоритмы, и, как результат, отсутствие универсальности, т. е. необходимость под каждый алгоритм изобретать уникальное устройство. Компьютеры — ярчайший пример огромного эффекта от фундаментальных научных исследований. А ведь для подавляющего большинства людей эти математические исследования кажутся полностью оторванными от жизни. Вряд ли кто-то мог предполагать, что исследование глубоких проблем в основах математики, выполненное несколькими учеными, даст результат, который за полвека полностью изменит облик цивилизации. Это изменение осуществляется не столько напрямую — через бытовое использование компьютеров, сколько косвенно — через качественное изменение научных исследований. Ведь многие исследования практически во всех отраслях науки требуют обработки огромных массивов данных, анализ которых для человека является непосильной задачей. Современная физика, астрономия, генетика или фармакология немыслимы без компьютеров.

Вдумайтесь: теоретическая концепция компьютеров была предложена до их физического воплощения, и было совершенно не ясно, для каких именно программ компьютеры могут предназначаться. Однако концепция универсальных машин была настолько мощной, что осталась неизменной в современных компьютерах и оказалась применимой для выполнения любых программ — от сложения двух чисел до игр, офисных приложений и различных систем автоматического управления.

Поскольку универсальная машина может эмулировать любую машину Тьюринга, очевидно, она может эмулировать и любую другую универсальную машину, и это позволяет считать алгоритмы независимыми от их аппаратной реализации. Действительно, один и тот же алгоритм можно

реализовать на обычном цифровом процессоре, с помощью оптических вычислений или даже на основе спичечных коробков (и такие реализации, если не универсальных машин, то отдельных алгоритмов, делались). Забегая вперед, можно также отметить, например, что искусственные нейронные сети могут быть рассмотрены в качестве еще одной формализации понятия алгоритма, и эта формализация не мощнее универсальной машины Тьюринга.

Все это дает повод думать, что проблема искусственного интеллекта — это, в первую очередь, алгоритмическая проблема. С такой точкой зрения связана сформулированная Ньюэллом и Саймоном *гипотеза физической символьной системы*, согласно которой для достижения интеллектуального поведения системой необходимо и достаточно, чтобы физическая система выполняла преобразование символьной информации. Как следствие, можно сказать, что реализация сильного ИИ возможна на любом физическом носителе.

Обратите внимание: компьютеры — физическое воплощение концепции алгоритма, которое было введено для формализации мышления в процессе решения математических задач. Именно благодаря этому компьютеры оказались столь универсальными. Тезис Чёрча—Тьюринга можно интерпретировать как утверждение, что человеческий мозг эквивалентен универсальной машине Тьюринга (снабженной, однако, большим количеством специализированных алгоритмов). Из этого, в частности, следует, что человеческое мышление можно будет эмулировать, коль скоро у нас появится его исчерпывающее описание, которое можно подать на вход универсальной машины Тьюринга.

Формализация понятия алгоритмов и последующее появление компьютеров сделали возможной науку об искусственном интеллекте, однако не позволили тут же его создать. Существующие по сей день трудности в моделировании интеллекта являются одной из причин, по которой многие люди считают мышление неалгоритмизуемым. Помимо неконкретных эмоциональных высказываний типа: «Машина может делать только то, что в нее заложил человек, а сам человек способен творить» (что в определенной мере является заблуждением относительно возможностей как машины,

так и человека) или «Машина принципиально не может испытывать чувств», для этого имеются и более весомые аргументы. Одним из аргументов против алгоритмизуемости мышления, требующих серьезного рассмотрения, является существование алгоритмически неразрешимых проблем, которые остаются не только в формальных системах Гёделя, но также и для машины Тьюринга и ее эквивалентов.

АЛГОРИТМИЧЕСКАЯ НЕРАЗРЕШИМОСТЬ

В замечательном произведении Аркадия и Бориса Стругацких «Понедельник начинается в субботу» есть такой диалог:

– Голубчики, — сказал Фёдор Симеонович озабоченно, разобравшись в почерках. — Это же проблема Бен Бецалеля. Калиостро же доказал, что она не имеет решения.

– Мы сами знаем, что она не имеет решения, — сказал Хунта, немедленно ощетиниваясь. — Мы хотим знать, как её решать.

– Как-то странно ты рассуждаешь, Кристо... Как же искать решение, когда его нет? Бессмыслица какая-то...

– Извини, Теодор, но это ты очень странно рассуждаешь. Бессмыслица — искать решение, если оно и так есть. Речь идёт о том, как поступать с задачей, которая решения не имеет. Это глубоко принципиальный вопрос, который, как я вижу, тебе, прикладнику, к сожалению, не доступен.

Решение неразрешимых задач — это не просто художественная метафора. Поспорить здесь можно лишь с тем, что этот вопрос принципиален не в меньшей степени и для практики. Как это ни парадоксально, вся область искусственного интеллекта посвящена, по сути, решению неразрешимых и плохо поставленных задач. Но что это за задачи?

Как уже отмечалось, первые выводы о неразрешимости некоторых математических проблем были получены Гёделем в 1931 году. Наиболее известными являются две теоремы Гёделя о неполноте формальной арифметики. Их сущность в следующем. Если мы возьмем набор непротиворечивых аксиом, описывающих свойства чисел и арифметических операций над ними, то будут существовать некоторые утвержде-

ния о числах, которые не могут быть ни доказаны, ни опровергнуты на основе выбранных аксиом (такие утверждения называются *невыводимыми*). Более того, некоторые из этих утверждений являются вполне естественными истинными (с точки зрения человека) утверждениями. Чтобы «доказать» эти утверждения, приходится вводить новые аксиомы. Но какую бы мы формальную систему ни взяли, всегда найдутся утверждения об объектах системы, истинность которых не может быть установлена в рамках самой системы. Человек же каким-то «мистическим» образом определяет, должны ли эти утверждения быть истинными или ложными, и может расширить формальную систему так, чтобы получить желаемый результат. Что, по-вашему, должно это значить? Некоторые мыслители делают вывод о невозможности формализации мышления, особенно его творческой составляющей. А поскольку машина Тьюринга — тоже формальная система, они полагают, будто мышление невозможно реализовать на компьютере.

Действительно, некоторые задачи не имеют полного алгоритмического решения. Наиболее широко известной является так называемая *проблема останова*. Задача останова заключается в том, что необходимо определить, останавливается ли некоторая программа при любых исходных данных или в некоторых случаях «зацикливается» (работает бесконечно долго). Проблема же заключается в том, что доказано отсутствие алгоритма, который бы мог решить задачу останова для любой программы. Иными словами, задача останова является алгоритмически неразрешимой. Важность проблемы останова связана с тем, что, как принято считать, некоторый алгоритм решил некоторую задачу, если он *остановился*, когда состояние ленты соответствует ответу. Если останова не происходит, то считается, что задача не решена, поскольку непонятно, в какой момент времени на ленте оказывается решение. Алгоритм, который решает задачу останова, тоже должен останавливаться, на чем и основывается доказательство неразрешимости проблемы останова. В этом доказательстве используется аналог парадокса лжеца. Парадокс возникает, если составить утверждение, отсылающее к самому себе. Например, является ли ложным следующее

высказывание лжеца: «Я всегда лгу, даже сейчас»? Как ту же хитрость использовать для проблемы останова?..

Рассмотрим идею доказательства неразрешимости проблемы останова от противного. Пусть она разрешима, т. е. существует алгоритм T, которому на вход можно подать некоторую программу, и он напечатает «0», если программа «зацикливается», и «1» — если останавливается. Рассмотрим такую программу P, которая просто вызывает алгоритм T, передавая на вход свое описание $T(P)$, и, если алгоритм T вернул «0», алгоритм P останавливается, а если алгоритм T вернул «1», алгоритм P преднамеренно «зацикливается», запуская бесконечный цикл. Для любого алгоритма T несложно составить программу P, которая, очевидно, существует. Но что будет, если эту программу запустить? Зациклится ли она или остановится? Если программа P останавливается, тогда алгоритм T, предположительно решающий проблему останова для любой программы, при вызове $T(P)$ должен вернуть «1», но тогда программа P должна зациклиться. Если программа P зацикливается, тогда алгоритм $T(P)$ должен вернуть «0», но тогда программа P должна была бы остановиться. Пришли к противоречию, то есть исходное допущение о существовании алгоритма T неверно.

Существует множество и других алгоритмически неразрешимых задач. В частности, алгоритмически неразрешимой является десятая проблема Гильберта о диофантовых уравнениях, доказательство чего в 1970 году было завершено советским математиком Юрием Владимировичем Матиясевичем. Оказывается, при наличии решения его можно получить за конечное (но неограниченное) число шагов, но если решения нет, в произвольном случае ни один алгоритм это установить не может.

Решения некоторых проблем до сих пор не найдены, и неизвестно, являются ли эти проблемы неразрешимыми. Иногда у них удивительно простые формулировки. Одной из старейших таких проблем является проблема Эйлера середины XVIII века: любое четное число не меньше четырех можно представить в виде суммы двух простых чисел ($4 = 2 + 2$,... $18 = 5 + 13$, ...). Эту проблему Эйлер сформулировал в ответ на гипотезу Гольдбаха, согласно которой любое

нечетное число не меньше семи можно представить в виде суммы трех простых чисел. Гипотеза Гольдбаха была не так давно доказана, тогда как проблема Эйлера до сих пор не решена. Вероятно, это не относится к проблеме Эйлера, но некоторые математические утверждения могут быть истинные, но недоказуемые (невыводимые, алгоритмически неразрешимые) в рамках данной аксиоматики, о чем и говорит теорема Гёделя.

Неужели возможности алгоритмов настолько ограничены, и на основе компьютеров можно создать только неуниверсальный, слабый ИИ? Значит ли это, что мышление действительно неалгоритмизуемо? Чтобы разобраться в этом, нужно учитывать ряд моментов.

Первый момент заключается в том, что алгоритмические проблемы обычно формулируются как массовые проблемы, т. е. проблемы, в которых требуется найти единый алгоритм решения бесконечной серии однотипных задач. Алгоритмически неразрешимые проблемы являются «слишком массовыми», например в проблеме останова требуется решить задачу *для любого алгоритма* и для любых исходных данных за конечное время. Очевидно, ни один человек не в состоянии решить задачу останова для любого алгоритма, иначе не было бы «зависающих» операционных систем и прочих программ (а ведь это еще относительно простые программы по сравнению с тем, какие невообразимо сложные программы находятся в множестве «всех алгоритмов»). Человек решает задачу останова, но делает это с ошибками, вероятность которых повышается с усложнением программ. Способность человека решать алгоритмически неразрешимые проблемы (как массовые проблемы) является крайне сомнительной. Его способность находить решения для отдельных частных случаев ничего не доказывает, ведь это под силу и компьютеру. А без этого пафосные заявления о принципиальной ограниченности компьютеров по сравнению с человеком становятся малосостоятельными.

Представьте себя в ситуации, аналогичной той, в которой оказывается программа T в нашем доказательстве. Пусть имеется автомат, который печатает «0», если вы говорите «1», и печатает «1», если вы говорите «0». И вас просят

ответить на вопрос, что напечатает автомат. При этом вам разрешено сказать только «0» или «1». Естественно, автомат напечатает противоположное тому, что вы сказали, и вы об этом знаете, но ничего сделать не можете. Не кажется ли, что такая «задача» была бы просто издевательством над вами?

У такого эксперимента есть интересная практическая реализация: дело в том, что осознание многих решений происходит несколько позже (иногда более чем на полсекунды), чем это решение возникает в мозгу. Простейшие решения (выбор из двух альтернатив, скажем, поднятие правой или левой руки) можно детектировать в виде потенциала готовности еще до того, как они оказываются осознанными. В действительности, момент осознания соответствует не моменту принятия решения; он «откладывается» до момента совершения самого действия — именно поэтому мы не замечаем задержки в выполнении телом осознанных команд (у больных с нарушенным механизмом задержки может возникать впечатление, что ими управляют извне). Подобные эксперименты с использованием электромиографов, фиксирующих мышечные биотоки, а позднее — электроэнцефалографов, «считывающих» некоторую информацию прямо с мозга, проводились начиная с 1970-х годов. Интересно, что человека просят самого выбирать, какую руку и в какой момент поднимать, оставляя ему полную свободу воли. И тем не менее, машина, получающая ЭЭГ-сигнал, узнает о решении испытуемого до него самого. Пусть одна рука человека означает «0», а другая — «1», и пусть машина печатает число, противоположное числу, выбранному человеком. Если испытуемому нужно будет угадать, что печатает машина, он никогда не сможет правильно выбрать ответ, хотя машина будет показывать ответ почти за секунду до того, как выбор человека будет возникать в его сознании. Эта проблема «человечески неразрешима».

А ведь алгоритм T находится именно в такой ситуации: с него считывается информация о выборе, как и с мозга в примере с человеком. Нельзя выбрать правильный ответ, когда правильность ответа меняется от самого выбора. Причем здесь мы требуем, чтобы алгоритм T обязательно напечатал «0» или «1», а в качестве входа у него выступала

только программа *P*. Какой-нибудь другой алгоритм, «видя», что анализируемая программа сама его вызывает, мог бы напечатать и что-нибудь другое, скажем, «42». Даже если бы алгоритм *T* был сверхинтеллектуальным и полностью понимал ситуацию, он мог бы разве что описать экспериментаторам все, что о них думает, но правильный ответ о зацикливании программы *P* дать не смог бы. И дело тут вряд ли в ограниченности понятия алгоритма. Действительно, если предположить, что создано некоторое неалгоритмическое устройство, решающее проблему останова, к нему можно применить приведенные выше рассуждения, просто заменив термин «алгоритм» другим термином, например «устройство неалгоритмического преобразования информации».

В действительности, такие гипотетические устройства Тьюрингом были рассмотрены еще в 1939 году. Это так называемые *машины с оракулом*. Под оракулом понимается некая сущность, способная «вычислять» невычислимые функции или решать алгоритмически неразрешимые проблемы. Тьюринг показал, что для таких машин проблемы, сформулированные относительно них самих (например, проблема останова машины с оракулом), ими же являются неразрешимыми. Это напоминает древний парадокс: может ли всемогущий бог создать камень, который сам не сможет поднять?.. Поэтому даже если сверхтьюринговые вычисления физически реализуемы, для них также найдутся неразрешимые проблемы. Видимо, что-то не так с самой формой постановки этих проблем, неразрешимых ни алгоритмически, ни божественно.

Следует иметь в виду, что доказательство алгоритмической неразрешимости некоторой задачи зачастую заключается в сведении задачи к проблеме останова, которая неразрешима просто в силу того, что любой фиксированный алгоритм не может быть сложнее самого себя. И тем не менее, факт существования «алгоритмически неразрешимых» проблем остается; вопрос лишь в том, как его интерпретировать.

Вторым существенным моментом является то, что компьютерные программы, хотя и имеют форму *алгоритмических процессов*, не являются формальными алгоритмами. Есть важное отличие реального компьютера от машины

Тьюринга — это его взаимодействие с внешним миром. В машине Тьюринга содержимое ленты фиксировано и меняется только самой машиной. В компьютер могут поступать новые данные и даже программы в процессе его работы. Как человек выходит за рамки формальных систем, например, откуда он знает, какие утверждения, невыводимые в формальной арифметике, считать истинными, а какие — ложными? Ответ прост — из опыта. Для человека «входная лента» — потенциально вся Вселенная (где на некотором языке «записаны» алгоритмы мышления, которые мозгом, как универсальной машиной, лишь исполняются). Когда мы затрагивали аксиоматический подход в математике, отмечалось, что выбор аксиом произволен, но интересные аксиомы выбираются так, чтобы описывать что-то в реальном мире. Дирак писал: «...те принципы, которые находит интересными математик, оказываются как раз теми, которые выбраны Природой». Но, по сути, математикам неоткуда брать свои аксиомы и принципы, кроме как из обобщения опыта. Более сложным является вопрос, каков алгоритм этого обобщения, и алгоритм ли это. Этот вопрос мы на время отложим.

Искусственный интеллект, реализованный на компьютере, будет ограниченным только в том случае, если он будет полностью изолирован от мира. Тогда он действительно будет представлять собой формальную систему, к которой применим результат теоремы Гёделя, и уметь только то, что мы в него заложили. Если же для физически реализованного алгоритмического процесса «входной лентой» будет также вся Вселенная, то для него указанная постановка проблемы останова будет просто некорректной: на входной ленте у алгоритма *T* кроме программы *P* будет еще и сам алгоритм *T*, и многое другое. Кроме того, программа *P* не сможет обратиться к работающей копии алгоритма *T* и только лишь вызовет другую его копию, тем самым парадокс будет снят. В этом случае выводы из неразрешимости проблемы останова, равно как выводы из теоремы Гёделя, к такой программе будут просто неприменимы. Хотя это вовсе не доказывает алгоритмическую разрешимость проблемы останова, но устраняет разницу между компьютером и человеком, когда они поставлены в одинаковые условия. Здесь

также уместно вспомнить центральную идею концепции автоматного программирования профессора А. А. Шалыто: машина Тьюринга, лишенная ленты, — это очень простое устройство, *конечный автомат*, поэтому даже поведение конечного автомата может быть столь же сложным, как поведение машины Тьюринга, если этот автомат будет использовать внешний мир как ленту.

Третий момент заключается в требовании к останову алгоритма. Конечно, требовать от алгоритмов остановки с выдачей результата вполне естественно для классических алгоритмов, однако нелепо требовать от искусственного интеллекта того, чтобы он спустя ограниченное время останавливался и выдавал какой-то конечный результат. Безостановочные алгоритмы обладают большей мощностью, чем классические алгоритмы, и тоже могут быть вполне полезны. К примеру, такой «алгоритм» может выдать какой-то результат и продолжить работать. Многим ученым не нравится идея безостановочных алгоритмов, поскольку для них остается вопрос, когда считать, что ответ сформирован, поэтому свойства таких алгоритмов в математике исследованы меньше. Но для них можно ввести понятие «вычислимости в пределе»: если на каком-то шаге формируется ответ, который не меняется в процессе последующей работы безостановочной программы, то полагается, что задача решена. К сожалению, мало известен тот факт, что классическая проблема останова является разрешимой в пределе. Таким образом, безостановочность алгоритма искусственного интеллекта (если таковой может быть построен) — это не плохое, а принципиально необходимое его свойство. Дело в том, что сложность безостановочных алгоритмов может неограниченно возрастать, в то время как алгоритмы, от которых требуется остановка за заранее определенное (в самом алгоритме или исходных данных) число шагов, действительно обладают ограниченными возможностями.

Итак, существование алгоритмически неразрешимых проблем и ограничения, накладываемые теоремой Гёделя на формальные системы, по-видимому, не являются бесспорным аргументом против возможности создания сильного искусственного интеллекта как физической реализации некоторо-

го алгоритмического процесса, но процесса безостановочного и открытого.

Имеются и другие «аргументы» против алгоритмизуемости мышления. К примеру, часто говорится, что результат работы алгоритма полностью и однозначно определяется начальными данными, в то время как человек обладает «свободой воли». Однако детерминированность не является ключевым свойством алгоритма. Существуют вероятностные и недетерминированные машины Тьюринга, результат работы которых неоднозначен. Однако установлено, что они не расширяют понятия алгоритма в смысле алгоритмической разрешимости задач. И хотя некий источник хаоса может быть важен для реализации мышления, идею алгоритмичности это не опровергает. Кроме того, существенной неопределенностью обладают данные, поступающие в программу из внешнего мира.

Интереснее другое: механическое применение известного алгоритма для нас не выглядит проявлением умственной деятельности. Если мы не знаем алгоритма умножения двух чисел «в столбик», то умножение каждой конкретной пары чисел для нас будет требовать интеллектуального напряжения. Но как только алгоритм известен, умножение любых двух чисел становится малоинтеллектуальной задачей. При обсуждении истоков возникновения понятия алгоритма мы видели, что алгоритм был нужен для описания воспроизводимого процесса доказательства теорем, но был также задан вопрос: откуда берутся сами эти доказательства? Может, мышление — это процесс построения новых алгоритмов в случаях, когда способ решения некоторой массовой проблемы еще неизвестен? И алгоритмичен ли этот процесс?

Поиск алгоритма, решающего проблему останова и ряда других проблем, в общем случае невыполним. Но это не мешает исследовать алгоритмы решения этих задач при некоторых естественных ограничениях. Зачастую поиск этих ограничений и составляет основную часть решения реальной «неразрешимой» задачи. Не зря говорят, что правильная постановка задачи составляет уже половину ее решения. Можно ожидать, что упрощение некоторой неразрешимой задачи сделает ее разрешимой, но весьма сложной. Также

и для многих других задач алгоритмы их полного решения найти можно, однако такой поиск обладает высокой сложностью. Не только задачи построения алгоритмов требуют огромного числа операций. Существует большой класс задач, для которых отсутствуют «быстрые» алгоритмы. Как правило, такие задачи производят впечатление интеллектуальных. Наиболее типичный пример — шахматы. Тесная связь этих задач с проблематикой мышления требует их отдельного обсуждения.

ПРОКЛЯТИЕ РАЗМЕРНОСТИ

Один алгоритм может применяться к разным входным данным, описывающим условие индивидуальной задачи из некоторой серии однотипных задач. При этом в зависимости от длины входных данных может меняться число операций, которые совершает алгоритм для нахождения решения.

Рассмотрим простой пример алгоритма, переводящего запись числа M из двоичной системы счисления в шестнадцатеричную. Этот алгоритм будет брать блоки по четыре бинарных символа и определять соответствующую им цифру в шестнадцатеричной системе. Очевидно, время работы такого алгоритма будет пропорционально количеству двоичных разрядов в исходном числе, т. е. пропорционально длине входных данных (количество разрядов в числе M можно оценить как $N = [\log_2 M]$).

Рассмотрим теперь алгоритм, решающий другую задачу: установить, является ли заданное число M простым. Алгоритм непосредственной проверки простоты будет перебирать все числа от 2 до $\sqrt{M}$ и смотреть, делится ли число M нацело на какое-либо из них. Длина входной строки здесь также будет $N = [\log_2 M]$, а вот время работы в худшем случае будет пропорционально $\sqrt{M} \propto \sqrt{2^N} = 2^{N/2}$. Обратите внимание на то, что число действий определяется в зависимости от длины входной строки N, а не от числа M. Длина строки (количество информации) является более универсальной характеристикой, поскольку она может быть посчитана, даже если входные данные имеют нечисловой характер.

Зависимость числа операций, выполняемых алгоритмом, от длины входной строки называется *вычислительной сложностью* алгоритма. Понятно, что два рассмотренных алгоритма имеют разную вычислительную сложность.

Пусть мы запускаем оба алгоритма на компьютере, выполняющем миллиард операций в секунду. И пусть первый алгоритм сопровождается сложной графикой, требующей для отображения каждого числа 5 млн операций, а второй алгоритм проверяет делимость числа M на любое другое число за одну операцию. Если алгоритм работает меньше 1/10 с, то считаем, что он работает почти мгновенно. В таблице приведено, как будет меняться время работы каждого алгоритма при разных длинах входных данных. Для сравнения нижней строкой в таблице приведено время работы второго алгоритма на вычислительной системе с производительностью в тысячу раз больше.

Оценка времени работы алгоритма проверки простоты числа

M	Число разрядов N						
	10	20	40	60	80	100	120
	Время работы алгоритма						
$5 \cdot 10^6 N$/1 ГГц	<0,1 с	0,1 с	0,2 с	0,3 с	0,4 с	0,5 с	0,6 с
$2^{N/2}$/1 ГГц	<0,1 с	<0,1 с	<0,1 с	0,1 с	1100 с	13 дней	37 лет
$2^{N/2}$/1 ТГц	<0,1 с	<0,1 с	<0,1 с	<0,1 с	1,1 с	1126 с	13 дней

Видно, что время работы второго алгоритма после некоторого момента катастрофически возрастает, и даже существенное увеличение быстродействия компьютера не позволяет решить эту проблему: дальнейшее незначительное увеличение длины входной строки снова делает алгоритм невыполняемым за обозримое время.

Все алгоритмы разделяют на *классы сложности* в зависимости от характера возрастания числа выполняемых ими операций при увеличении длины входных данных N. К *классу P* относят алгоритмы, время которых возрастает не

быстрее, чем некоторый многочлен (polynomial) от N. Частный случай полиномиальной зависимости — это линейная зависимость, пример которой мы уже видели. Алгоритмы класса P считаются быстрыми в том смысле, что могут применяться на практике. Существуют разнообразные хорошо известные алгоритмы класса P, выполняющие сортировку массивов, поиск некоторой подстроки в строке, решающие системы линейных уравнений и т. д.

Второй рассмотренный нами алгоритм не относится к классу P: для любого наперед заданного многочлена найдется такое значение N, что величина $2^{N/2}$ окажется больше значения этого многочлена при том же N. Алгоритмы со временем работы, возрастающим быстрее любого полинома, помещаются в класс NP. Можно подумать, что аббревиатура NP происходит от «not polynomial», но на самом деле она означает «nondeterministic polynomial». Дело в том, что эти алгоритмы исполняются за полиномиальное время на недетерминированных машинах Тьюринга. Эти (гипотетические) машины вскользь уже упоминались. Их сущность заключается в том, что на каждом шаге машина может переходить не в одно, а сразу в несколько состояний, находясь в них как бы одновременно.

Ранее говорилось о том, что универсальная машина Тьюринга может эмулировать любую другую машину с полиномиальной скоростью. Теперь видно, что это важно в том смысле, что класс сложности исполняемого алгоритма при такой эмуляции не меняется.

Любой алгоритм класса P принадлежит также и классу NP (поскольку на недетерминированной машине он также будет выполняться за полиномиальное время). Но для простоты можно считать, что алгоритмы класса NP — это алгоритмы, время работы которых возрастает по размерности задач N экспоненциально (т. е. в геометрической прогрессии).

Таких алгоритмов тоже очень много. К сожалению, плохие вычислительные свойства относятся не только к алгоритмам, но и к самим задачам. Задаче тоже можно приписать некую вычислительную сложность, которую следует определить как сложность алгоритма с самым медленным возрастанием числа операций с ростом длины входных данных, называющейся также *размерностью задачи*.

Существуют задачи класса NP, для которых неизвестно алгоритмов класса P. Такие задачи мы будем называть *NP-полными* (хотя, строго говоря, некоторые задачи относят к промежуточным классам сложности). Установлена сводимость многих NP-полных задач друг к другу. Если для какой-то из них будет найден быстрый алгоритм, то на его основе можно будет построить быстрые алгоритмы решения и всех остальных задач. До сих пор нет доказательства того, что классы P и NP различны, т. е. что NP-полные задачи не могут решаться полиномиальными алгоритмами. Эта проблема, поставленная независимо Л. Левиным и С. Куком в 1971 году, входит в список «Millenium Prize Problems», состоящий из семи математических проблем, за решение каждой из которых Математическим институтом Клэя назначена премия в 1 млн долл. Многие практики, работающие с конкретными NP-полными задачами, убеждены, что P- и NP-классы не совпадают. Многие математики, пытавшиеся решить эту проблему, уверены, что сама она является неразрешимой (что, однако, тоже пока не доказано). И лишь оптимисты уверены в эквивалентности этих классов.

Почему же этому вопросу уделяется такое внимание? Время решения NP-полной задачи возрастает экспоненциально с ростом размерности задачи, поэтому такую ситуацию называют *комбинаторным взрывом*, пример которого мы видели в приведенной выше таблице. Это название связано с тем, что число многих комбинаторных объектов, таких как битовые строки или сочетания элементов множества, растет экспоненциально с увеличением их длины. Здесь уместно вспомнить о легенде, согласно которой мудрец, придумавший шахматы, попросил у индусского царя в качестве награды положить ему на доску зерна пшеницы: на первую клетку одно, а на каждую последующую — в два раза больше, чем на предыдущую. По легенде царь поначалу был рассержен пожеланием столь малой награды, но сейчас легко подсчитать, что число зерен было бы почти 10^{20}, и если бы каждая последующая клетка заполнялась через секунду, то это выглядело бы как гигантский взрыв. Стоит отметить, что и в настоящее время на вопрос о том, какова будет толщина листа бумаги, если его сложить пополам 50 раз подряд, значительная часть людей

отвечает, что толщина составит до нескольких метров (хотя, опять же, несложно ее оценить примерно как расстояние от Земли до Солнца). Также быстро растет и число требуемых операций для решения NP-полных задач.

Эти задачи, обычно сводящиеся к перебору часто встречающихся комбинаторных объектов, оказываются практически неразрешимыми при достаточно большом N, как бы быстродействие компьютеров ни возрастало и как бы ученые ни пытались придумать для них полиномиальные алгоритмы. При этом NP-полные задачи являются гораздо менее надуманными, чем алгоритмически неразрешимые проблемы. Исследователи образно охарактеризовали данную ситуацию как *проклятие размерности*.

Алгоритмы класса NP, как правило, не могут применяться на практике, если только объем исходных данных не является сильно ограниченным. В частности, на этом факте основаны многие методы криптографии. К примеру, пусть мы знаем два больших простых числа. Мы можем их перемножить и результат перемножения сделать открытым ключом, с помощью которого происходит шифрование, а сами простые числа сделать закрытым ключом, с помощью которого происходит дешифрование. Мы можем передать открытый ключ по незащищенному каналу связи, получить по этому же каналу зашифрованную информацию и дешифровать ее с использованием известного нам закрытого ключа. Хотя нет защищенной передачи информации, способ практически безопасен (при достаточной длине ключа), поскольку алгоритм шифрования относится к классу P, а для дешифрования потребуется алгоритм класса NP (конечно, здесь приводится лишь общая идея, а не описываются сами алгоритмы шифрования). Еще остается вопрос, откуда нам самим взять большие простые числа. Из-за сложности получения таких чисел и эффективности их использования для шифрования они составляют предмет охраняемой тайны.

Кстати, здесь виден один интересный момент, связанный с NP-полными задачами. Для них проверка правильности единичного решения обычно является весьма простой (например, деление двух чисел). И даже сама задача класса

NP может быть определена как задача, единичное решение которой может быть проверено за полиномиальное время на машине Тьюринга (это определение считается эквивалентным определению через недетерминированные машины). Но поиск решения задачи сводится к перебору и проверке очень большого числа вариантов. Алгоритмы же класса P (применение которых не вызывает ощущения интеллектуальной деятельности), как правило, решение вычисляют в явном виде, без перебора.

Большинство задач в области искусственного интеллекта, если и не являются алгоритмически неразрешимыми, то, по крайней мере, являются NP-полными (строго говоря, есть разрешимые задачи, которые сложнее, чем NP-полные задачи, но они не так сильно распространены). Такие NP-полные задачи, как игра в шахматы или построение доказательства теорем, кажутся интеллектуальными. Но алгоритмы точного решения этих задач, реализуемые на компьютерах (воплощениях машины Тьюринга), являются неприменимыми на практике. Значит ли это в очередной раз, что искусственный интеллект не может быть реализован на компьютерах? Посмотрим, есть ли какие-то альтернативы, чтобы спастись от комбинаторного взрыва.

ПАРАЛЛЕЛЬНОСТЬ, НЕПРЕРЫВНОСТЬ И КВАНТОВЫЕ КОМПЬЮТЕРЫ

Если программно преодолеть проклятие размерности нельзя, может, существует возможность аппаратного решения? Наиболее очевидной является идея физической реализации недетерминированной машины Тьюринга (на которой NP-полные задачи решаются за полиномиальное время) с помощью параллельных вычислений: все состояния, в которых одновременно находится недетерминированная машина, могут просто просчитываться на отдельных процессорах.

Действительно, в современных цифровых компьютерах общего назначения вычислениями занимается только процессор с ограниченным набором регистров и весьма неболь-

шим кэшем. Как и в машине Тьюринга, имеется указатель на текущую инструкцию, которая и выполняется. В грубом приближении можно сказать, что используется лишь один вычислительный элемент (имеющий, правда, уже весьма сложную внутреннюю организацию), через который последовательно пропускаются команды. Это обеспечивает однозначность выполнения и простоту представления алгоритмов. Сейчас процессоры стали развиваться по пути увеличения числа отдельных «вычислителей» — ядер, но их количество остается незначительным. Память компьютера состоит из огромного числа элементов, которые, однако, не выполняют операций по преобразованию информации.

Такая организация вычислений разительно отличается от организации человеческого мозга, в котором каждый нейрон является отдельным вычислителем, работающим параллельно с другими. Можно представить мощность компьютера, у которого каждая ячейка памяти не просто хранит какую-то информацию, но и способна ее самостоятельно обрабатывать, обращаясь к информации в других ячейках! Может, мощность человеческого мышления кроется в параллельной обработке информации? Но ведь по грубым оценкам человеческий мозг способен выполнять порядка 10^{14} операций в секунду (поскольку он состоит из 10^{11} нейронов, каждый из которых передает до 1000 импульсов в секунду), что сопоставимо с мощностью современных компьютеров. Существуют, правда, гипотезы, что вычислительная мощность мозга гораздо больше благодаря сложным процессам, идущим внутри самих нейронов (одноклеточные организмы обладают сложным поведением, которое обеспечивается внутренней системой управления). Однако в мозге во многих группах нейронов одновременно активируется малый процент нейронов, и это отражается в том, что на уровне сознания мышление человека достаточно последовательное.

В настоящее время существуют и цифровые компьютеры, реализующие массовые параллельные вычисления. Это программируемые логические интегральные схемы (типа FPGA — Field-Programmable Gate Array), состоящие из большого числа логических блоков, между которыми могут устанавливаться различные связи. Хотя сами блоки являются

сравнительно маломощными, благодаря их параллельной работе общая производительность может оказаться выше, чем у обычных процессоров. Очень популярными сейчас становятся вычисления на графических процессорах общего назначения (GPGPU), которые могут быть в десятки и сотни раз эффективнее вычислений на CPU. Однако множество задач, для которых удается добиться заметного выигрыша, ограничено. Наиболее типичными приложениями, для которых использование таких схем является оправданным, являются задачи по обработке изображений (и других сигналов), где каждая точка изображения может обрабатываться одной и той же процедурой параллельно, а также аппаратная реализация искусственных нейронных сетей некоторых типов.

Другим интересным примером систем с массовой параллельностью являются кластеры компьютеров. Можно представить себе почти не фантастическую ситуацию, когда подобный кластер охватывает все компьютеры в Интернете, общее число которых не существенно меньше числа нейронов в мозге, а связность между компьютерами даже больше. В настоящее время существуют различные проекты, использующие возможности распределенных вычислений. К ним относится, например, проект SETI@home, в котором используются свободные вычислительные мощности на компьютерах миллиона добровольцев для обработки астрономических данных в целях поиска сигналов внеземных цивилизаций. Аналогичными характеристиками обладает проект Folding@home, в котором распределенные вычисления используются для моделирования процессов образования трехмерной структуры белков в рамках генетических исследований.

Хотя на практике системы с массовой параллельностью могут быть полезны, в целом они пока не дают принципиального выигрыша и не решают проблему NP-полноты. Чтобы такое решение возникало, необходимо, чтобы число элементов в системе росло экспоненциально с ростом размерности задачи. Даже миллиард процессоров, работающих параллельно над одной NP-полной задачей, легко «поставить в тупик», лишь немного увеличив ее размерность (например, если мы таблицу, приведенную в предыдущем разделе, расширим столбцом $N = 150$, то у миллиарда параллельно

работающих гигагерцевых процессоров уйдет 10,5 часов на ее решение..., а при $N = 300$ время решения будет намного превышать возраст Вселенной).

Некоторые ученые ищут пути преодоления проблемы комбинаторного взрыва (а то и алгоритмической неразрешимости) в отказе от дискретности. Действительно, возможности непрерывных (по сравнению с дискретными) процессов в принципе могли бы оказаться гораздо значительнее. Существуют попытки расширить понятие алгоритмов с дискретного на непрерывный случай. Ведь мощность множества (количество элементов в нем) всех вещественных чисел несопоставимо больше мощности множества целых чисел (в то же время множества целых и рациональных чисел равномощны).

В вычислительной технике, однако, наоборот, отказ от аналоговых в пользу цифровых систем дал существенные положительные результаты (этот факт, к примеру, подчеркивается в книге профессора В. В. Александрова «Цифровые технологии инфокоммуникации»). Аналоговые вычисления могли бы иметь принципиальное превосходство над цифровыми только в случае бесконечной точности их выполнения. Однако малейшие внешние возмущения делают точность конечной, причем ошибка в процессе вычислений накапливается, т. е. непрерывная система может быть эффективно заменена дискретной.

Дискретная аппроксимация непрерывного сигнала в известных системах не приводит к потере возможностей, а напротив, увеличивает надежность. Так, генетический код считается дискретным. Даже в мозге многое дискретно, например приходящий сигнал попадает на дискретные элементы — отдельные рецепторы, которые передают дальше информацию не непрерывно, а в виде отдельных импульсов — спайков. Если непрерывный характер процессов, обеспечивающих мышление, является таким уж необходимым, может показаться немного странным, почему тогда язык состоит из дискретных элементов. Почему мы составляем предложения из отдельных звуков и слов? Ведь сама речь передается как непрерывный сигнал, и можно было бы представить себе язык, в котором посредством речи передавались

бы целые образы (по некоторым сведениям, таким языком пользуются дельфины). Если полагать, что мышление обеспечивается какими-то непрерывными или невычислимыми процессами, использование мозгом непрерывного звукового сигнала для передачи сообщений, состоящих из дискретных элементов, окажется в высшей мере неестественным. Считается, что последовательной обработкой дискретной, в частности лингвистической, информации занимается левое полушарие, а правому полушарию мозга свойственны непрерывность и параллельность работы, однако это разделение идет, скорее, не на физическом, а на информационном уровне.

Хотя сейчас существуют интересные аналоговые решения, например в области оптических вычислений, эффект от их использования сродни эффекту от дискретных систем с массовым параллелизмом. Даже вопрос о дискретности самих физических процессов не до конца ясен. С одной стороны, почти все физические поля квантуются, т. е., истинно непрерывные вычисления, возможно, физически не реализуемы. С другой стороны, элементарные частицы представляются непрерывными волновыми функциями. Квантовый мир имеет и другие интересные свойства.

Несомненный интерес представляет преодоление проклятия размерности посредством «честного» воплощения недетерминированной машины Тьюринга в форме так называемых *квантовых компьютеров*, основа которых была заложена Ричардом Фейманом и Дэвидом Дойчем в 1980-х годах с использованием более ранних результатов других ученых.

Основной структурной единицей квантовых компьютеров является *кубит* — квантово-механическая система (например, отдельный атом), которая может находиться в одном из двух состояний. На первый взгляд, кубит выглядит как обычный классический бит памяти. Однако каждый кубит является активным элементом — он «эволюционирует» в соответствии с законами квантовой механики, переходя в другое состояние, что можно представить как вычисление. Квантовый компьютер может состоять из многих кубитов, эволюция каждого из которых будет зависеть от состояния прочих кубитов.

Казалось бы, такая совокупность кубитов представляет собой просто систему с массовой параллельностью. Но привлекательность квантовых компьютеров вовсе не в том, что они могут состоять из очень большого числа параллельно работающих элементов (это как раз обеспечить крайне сложно), а в том, что квантово-механическая система, по современным представлениям, обладает совершенно изумительным свойством — она может находиться в нескольких состояниях одновременно. Это именно то свойство, какое требуется от недетерминированной машины Тьюринга.

Но что это означает? Если классическая система из трех бит может находиться лишь в восьми состояниях: 000, 001, 010, ..., 111, то квантово-механическая система может находиться не только в этих восьми чистых состояниях, но также и в любых смешанных состояниях вида $|001\rangle + |100\rangle$ или $|000\rangle + |101\rangle + |001\rangle$, т. е. в их суперпозиции (здесь используются классические обозначения состояний в квантовой механике, но опущены нормировочные коэффициенты). При этом состояния разных кубитов могут быть «зацепленными» (или «спутанными»), т. е. не просто каждый кубит по отдельности (независимо от других) находится в состояниях $|0\rangle$ и $|1\rangle$, но вся система находится сразу в состояниях (для последнего примера) $|000\rangle$, и $|101\rangle$ $|001\rangle$. Эволюция каждого состояния соответствует вычислениям над собственным набором данных, и эти вычисления выполняются параллельно.

Многие воздействия на квантово-механическую систему переводят ее из одного состояния в несколько состояний, существующих одновременно. Например, единичный фотон при отражении от полупрозрачного стекла одновременно отразится от него и полетит насквозь, т. е. будет находиться сразу в обоих состояниях. И лишь при его регистрации произойдет таинственная редукция, в результате которой будет случайно выбрано одно из двух состояний.

Таким образом, квантовый компьютер действительно может быть истинным воплощением недетерминированной машины Тьюринга. Уже придуманы квантовые компьютеры, которые могут за полиномиальное время решать некоторые задачи, не относящиеся к классу P, например задачу о разделении числа на простые сомножители. Несмотря даже на то, что каждый такой компьютер предназначен для реализации одного алгоритма (хотя сейчас предлагают и перепрограммируемые компьютеры), этот результат производит весьма сильное впечатление — настолько неожиданным был приход из физики решения некоторых проблем информатики. Действительно ли квантовые компьютеры могут решить проблему NP-полноты?

Пока ответить на этот вопрос сложно. Не все ученые даже согласны с тем, что квантово-механические системы действительно находятся в нескольких состояниях одновременно. Возможно, это лишь удобное математическое представление, а реальная система находится в одном состоянии, выбираемом как-то случайно (с точки зрения измерения состояния квантово-механической системы все именно так и выглядит). Хотя существуют эксперименты, свидетельствующие в пользу реальности квантовой суперпозиции, их интерпретация может оказаться и иной. Да и на хорошо известный парадокс кота Шрёдингера до сих пор нет единого ответа. Этот парадокс представляет собой мысленный эксперимент: поместим в закрытую коробку кота и механизм, который при детектировании распада ядра радиоактивного элемента (также находящегося в коробке) убивает кота. Пусть вероятность распада ядра в течение часа — 50 %, точнее, через час ядро будет находиться в суперпозиции двух равновероятных состояний — распавшемся и не распавшемся. Если эти состояния существуют одновременно, то и кот будет одновременно и живым, и мертвым. Но если коробку открыть, можно будет увидеть лишь одно состояние кота. Возникает вопрос: когда именно вместо двух состояний появляется одно? Существуют разные мнения: и то, что нахождение даже атома в нескольких состояниях одновременно — абстракция, и то, что редукция суперпозиции происходит при наблюдении или измерении (хотя понятие наблюдения остается смутным), и то, что при открытии коробки наблюдатель сам переходит в суперпозицию существующих одновременно, но никак не взаимодействующих состояний, в одном из которых он видит мертвого кота, а в другом — живого (по сути, вся Вселенная трактуется как множество таких параллельных разветвляю-

щихся миров). Интересна интерпретация Пенроуза, согласно которой редукция суперпозиции тем вероятнее, чем больше разница энергий альтернативных состояний (что гораздо яснее идеи наблюдения). Окончательно реальность квантовой суперпозиции станет ясна лишь после создания настоящих квантовых компьютеров. Современные квантовые компьютеры не эквивалентны недетерминированным машинам Тьюринга.

Кроме того, даже если все теоретические предположения верны, может оказаться так, что на практике реализуемы лишь очень простые квантовые компьютеры, поскольку большие квантово-механические системы будет не только проблематично построить, но и крайне сложно оградить от возмущения со стороны окружения, нарушающего их работу.

Трудно сказать, насколько существенным может оказаться использование квантовой механики для реализации сильного искусственного интеллекта. Некоторые ученые выступают решительно против возможности воспроизведения мышления на обычных цифровых компьютерах, опираясь на существование алгоритмически неразрешимых проблем и проклятие размерности. Еще до появления компьютеров, когда о мозге было известно очень мало, исследователи пытались установить «химическую формулу» мысли или свести познавательные процессы к некоторым физическим законам (например, движению каких-то частиц в поле сил). Высказывание «мозг вырабатывает мысль как печень — желчь» стало саркастическим заметно позднее. И сейчас остались ученые, ищущие какие-то особые физические условия, необходимые для возникновения мышления. При этом квантовая механика с ее загадочными законами является первым кандидатом на эту роль.

Одним из наиболее заметных представителей этой позиции является упоминавшийся выше физик с мировым именем — Роджер Пенроуз. Однако, если в более ранней своей книге «Новый ум короля» он отстаивал точку зрения, что в мозгу могут происходить квантовые вычисления при решении NP-полных задач (а в качестве кубитов, по сути, выступают нейроны), то впоследствии (в книге «Тени разума») он модифицировал свою точку зрения. Ведь работа некоторых ансамблей нейронов неплохо моделируется на компьютере, а также имеются определенные успехи в моделировании целиком нервной системы простых организмов, таких как черви нематоды, в распоряжении которых имеется всего 302 нейрона, расположение и связи между которыми хорошо изучены. Даже отклики нейронов в зрительной системе обезьяны при предъявлении ей определенных стимулов воспроизводятся в моделях достаточно точно. Признав возможность алгоритмического описания процессов передачи сигналов между нейронами, Пенроуз предположил, что квантовые эффекты могут играть существенную роль внутри нейронов (в их цитоскелете), в частности, в процессе образования и обновления самих связей между нейронами. Причем эффекты эти, по его предположению, носят «невычислимый» характер и их следует искать за пределами современной физики среди еще неоткрытых законов.

Существование невычислимых физических процессов является красивой научной гипотезой. Представить саму возможность таких законов не так сложно: достаточно допустить, что, например, квантово-механическое состояние элементарной частицы изменяется не случайно, а в соответствии с решением некоторой алгоритмически неразрешимой задачи (например, диофантова уравнения, параметры которого определяются окружением). Еще совершенно не ясно, какие физические механизмы могут давать это решение (вера в их существование берется лишь из кажущейся неалгоритмичности человеческого разума). Пока некоторые физики ищут возможность обойти алгоритмичность, другим ученым остается работать в рамках классической алгоритмической парадигмы. Ведь даже если согласиться с утверждением Пенроуза о важности для мышления каких-то тонких физических процессов внутри клеток, никто не отменит тот простой факт, что человеческий мозг отличается от нервной системы, скажем, гидры, не протекающими в нем физическими процессами, а своей структурой, поддающейся алгоритмическому описанию. Кроме того, как уже отмечалось, проблема останова будет неразрешимой и для класса «неалгоритмических устройств».

Итак, для ряда ученых существование алгоритмически неразрешимых проблем и проклятие размерности служат

свидетельством в пользу невозможности реализации сильного ИИ на цифровых компьютерах. Но, может, тот факт, что человек не способен решать алгоритмически неразрешимые проблемы (например, не способен решить любое диофантово уравнение) и не способен устанавливать точные решения NP-полных задач за ограниченное время (например, не способен идеально играть в шахматы), служит веским доводом в пользу того, что ограничения человеческого разума во многом сходны с ограничениями универсальной машины Тьюринга, и какие бы то ни было «невычислимые» или просто квантово-механические процессы не являются принципиальной компонентой мышления, даже если таковые процессы в мозгу происходят?

С тем, что алгоритмическая компонента является весьма значимой в работе разума, согласны даже ярые сторонники неалгоритмичности мышления. И, как мы видели, работа алгоритмов не зависит от аппаратной реализации. Простое повышение быстродействия процессоров или числа элементов некоторой системы вряд ли само по себе даст возникновение разума, как это иногда представляется в фантастике. Например, маловероятно, что Интернет уже является разумным или может стать таковым без специального программного обеспечения, распределенным образом реализующего когнитивные функции. Первооткрыватели области искусственного интеллекта вполне обоснованно выбрали алгоритмический подход, и выбор этот был вдохновлен не только появлением компьютеров, но некоторыми данными о мышлении человека и животных, существовавшими к середине прошлого века.

ИЗМЕРЕНИЕ ИНТЕЛЛЕКТА

Многим людям мышление кажется очень сложным и загадочным процессом, ведь человек способен решать самые разнообразные задачи, писать стихи, изобретать сложные машины и так далее... Те же, кто полагает, будто мышление является чем-то простым, не дают ответа об устройстве

мышления, достаточного для его искусственного воспроизведения.

Как подступиться к проблеме искусственного интеллекта, с чего начать исследование? Вполне естественным было обратиться к изучению уже существующего интеллекта, единственным примером которого, как многие считают, является интеллект человеческий. Действительно, именно интеллект является единственным явным отличием человека от животных, и пропасть эта кажется столь глубокой, что даже порой сложно представить, как человек мог произойти от животных в результате эволюции (как мы, однако, увидим в дальнейшем, пропасть между ними не столь большая). Что же составляет самую суть человеческого мышления, чем оно отличается от мышления животных?

Самое очевидное отличие человека от животных — использование языка. Связь языка с мышлением настолько очевидна, что даже в фантастической литературе способность к использованию языка традиционно считается одним из основных критериев наличия интеллекта у инопланетных рас. Да и многие люди мыслят как будто словами. Даже учеными-психологами часто высказывалось мнение, что речь играет в мышлении решающую роль в соответствии с «формулой»: мышление равно речь минус звук. Но откуда соответствующие фразы появляются в мозгу? Как человек способен понимать чужие высказывания? И вообще, зачем мышлению язык?

Изучение структуры языка самого по себе не дает прямых ответов на эти вопросы. Усвоение и использование языка становится возможным только благодаря тому, что в мозгу уже существуют какие-то механизмы мышления. Кроме того, мышление может протекать и без использования слов (например, при игре в шахматы) и даже вовсе без осознания мыслей. Действительно, в словесной форме через сознание проходят десятки байт в секунду, в то время как мозг целиком должен обрабатывать, видимо, в миллиарды раз больше информации. Не удивительно, что наделить компьютер способностью пользоваться языком без реализации неких более глубинных механизмов мышления не получилось. Но что было известно о человеческом интеллекте вообще к 1950-м годам — к моменту возникновения области ИИ?

Почти до конца XIX века разум был предметом преимущественно философского рассмотрения. И хотя философы сделали ряд интересных наблюдений, их выводы слишком неконкретны, поскольку они опирались лишь на самонаблюдение (интроспекцию), без какого-либо объективного эксперимента. На практике же, например в психиатрии или педагогике, требовались теории мышления, на основе которых можно было бы разрабатывать методики обучения или лечения неврозов. Одной из первых содержательных теорий стал психоанализ Зигмунда Фрейда, больше ориентированный, правда, на патологические состояния психики и не содержащий каких-либо количественных оценок. Успехи естественных наук взывали к необходимости введения количественных характеристик в психологии, в частности, в оценке интеллекта.

Введение количественной оценки интеллекта выглядит вполне естественным: младенец вряд ли обладает интеллектом (в смысле способности решения произвольных задач), и в процессе развития ребенка дискретного перехода между «неинтеллектуальным» и «интеллектуальным» состоянием не наблюдается, т. е. уровень интеллекта должен увеличиваться непрерывно, а значит, этот уровень можно измерить количественно.

Первые тесты для выявления уровня интеллекта были предложены в 1883 году Гальтоном в монографии «Исследование человеческих способностей и их развитие». Они сводились к определению скорости реакции, способности определять характеристики зрительных и звуковых стимулов, например высоту тона.

Существенное развитие этих идей произошло в 1905 году, когда во Франции власти озаботились проблемой детей, отстающих в умственном развитии. Психологам была поставлена практическая задача выявления таких детей дошкольного возраста для их обучения по специальным программам.

А. Бине и Т. Симон разработали системы тестов для детей разных возрастов. Тесты для каждого возраста составлялись таким образом, чтобы в среднем дети меньшего возраста не могли бы на них отвечать. С помощью таких тестов можно было определять «умственный возраст», который мог отли-

чаться от биологического возраста, что служило индикатором умственной отсталости или одаренности.

Понятие коэффициента интеллекта (intelligence quotient — IQ) ввел В. Штерн как отношение умственного возраста к биологическому. Сам Бине был против интерпретации результатов тестирования как уровня интеллекта, полагая, что интеллект нельзя рассматривать как простую сумму элементарных способностей. Действительно, исходно IQ подразумевал не уровень интеллекта вообще, а степень сформированности некоторых умственных навыков. Однако удобство использования тестирования и кажущаяся объективность результатов привели к существенному расширению сферы применения подобных тестов.

Развитие психологии интеллекта, выделившейся в самостоятельную ветвь исследований, оказалось под сильным влиянием тестологического подхода. Тесты дифференцировались, разделялись по «первичным умственным способностям» на группы: тесты на пространственное мышление, на восприятие, на речевые способности, на арифметические операции, на выявление закономерностей, на память... Возник вопрос, есть ли «общий интеллект», или «первичные умственные способности» не зависят друг от друга? Ясный ответ на этот вопрос получен не был, поскольку, с одной стороны, люди, хорошо выполняющие одни тесты, часто неплохо справляются и с другими тестами, но, с другой стороны, связь эта далеко не однозначна и бывает даже обратной. Постепенно список первичных умственных способностей уточнялся, например в него была добавлена способность к обучению, однако ясности не то, что в механизмы работы мышления, но даже в его основные компоненты, это привнесло немного. Нельзя даже сказать, корректно ли по аналогии со спортом выделять интеллектуальных спринтеров и стайеров, или же в случае с мышлением способность к быстрому решению простых задач не входит в противоречие с медленным осмыслением глобальных проблем.

Корректна ли метафора «мышцы интеллекта», т. е. можно ли связывать низкий уровень интеллекта или отсутствие одаренности с плохой памятью, низкой скоростью чтения, малым словарным запасом, неспособностью рифмовать слова

или выполнять в уме арифметические операции? Можно ли интеллект рассматривать как простую сумму некоторых базовых навыков? В истории известно много случаев, когда человек достигал выдающихся успехов на пути преодоления, например косноязычия, становясь великолепным оратором, или слабого слуха, становясь знаменитым музыкантом. Часто человек, что-то хорошо умеющий от природы, не будет даже задумываться о том, как у него это получается, и не будет развивать свое умение. Напротив, человек с врожденным «дефектом» будет пытаться понять суть соответствующей деятельности и вполне может сверхкомпенсировать свой дефект.

Более того, некоторые дефекты могут способствовать развитию иных интеллектуальных навыков. Так, ученик с дислексией (нарушением способности чтения) часто расценивается как отстающий, и многие тесты такой ученик будет выполнять гораздо медленнее. Однако дислексия может быть вызвана лишь нарушениями некоторых специализированных участков зрительной системы и может не затрагивать каких-либо иных умственных способностей. Напротив, она может помогать более глубокому обдумыванию прочитанного. А ведь на обдумывание книг, которые действительно заслуживают прочтения, должно тратиться гораздо больше времени, чем на само прочтение. Конечно, приятно иметь способность за пять минут прочитать книгу и полностью запомнить ее содержание, но без понимания этого содержания смысла в таком чтении немного. Не так уж удивительно, что среди выдающихся ученых или режиссеров нередко встречаются дислектики, к которым нередко относят, к примеру, и Альберта Эйнштейна (насчет IQ которого существуют различные мифы: и то, что он был сверхвысоким — больше 200, и то, что он был просто высоким — 160, и то, что он был около или даже ниже среднего –100; в любом случае маловероятно, что Эйнштейн тестировался именно по современной шкале, поскольку умер раньше ее распространения).

Количественное измерение внешнего проявления умственной деятельности в очень упрощенных условиях, каковыми являются тесты, оказалось настолько бессильным в познании природы мышления, что сторонники тестологического

подхода стали даже говорить о том, что такой вещи, как «интеллект», просто нет в природе. С помощью факторного анализа (о котором мы поговорим заметно позднее) выделялись независимые «интеллекты» — вербальный, математический, пространственный, мнемонический и т.д. Некоторые ученые выделяли до 150 факторов, которым не было названий. Другие ученые, такие как Г. Айзенк, пошли на хитрость, введя понятие «психометрического интеллекта» — того свойства человека, которое может быть измерено с помощью тех или иных тестов, — не особо заботясь о том, как этот «психометрический интеллект» соотносится с реальным. Не случайно стала распространенной шутка, что подобные тесты измеряют лишь способность человека проходить сами эти тесты.

Существует даже международное сообщество под названием «Менса», в которое входит свыше 100 тысяч человек с очень высоким IQ. Среди них, однако, за редким исключением, отсутствуют люди, достигшие выдающихся результатов в какой-либо сфере деятельности. Напротив, многие нобелевские лауреаты, хотя и обладают относительно высоким IQ, в большинстве своем могли бы «не пройти по конкурсу» в это общество.

К сожалению, тестологический подход, несмотря на всю свою методологическую порочность (заключающуюся, в первую очередь, в исследовании внешних проявлений сложного феномена вместо изучения его внутренней структуры), проник даже в школьное обучение. Естественно, предварительное тестирование позволяет неплохо прогнозировать успешность обучения... ведь успешность обучения также оценивается по тестам.

Тем не менее в тестах IQ какая-то связь с интеллектом присутствует, и знакомство с ними может быть полезным. Посмотрим на типичные задачи в IQ-тестах.

1. Продолжите ряд: 2 3 7 13 27 ...
2. Вставьте пропущенное число:
196 (25) 324
329 (??) 137
3. Вставьте пропущенное слово:
РОТОР (РОСА) КАСКА

ГАРАЖ (. . . .) ТАБАК

4. Кто не является поэтом:

а) Кбол б) Икшнуп

в) Неисне г) Гшонантив

д) Воскарен

5. Вставьте пропущенное слово, являющееся концом первого и началом второго слов:

МЕ (. . .) ОЛАД

Интересно решить эти задачи, все же требующие какой-то элементарной умственной активности, и проследить процесс решения каждой их них.

Что же в этих задачах общего?.. Все они, хотя и ориентируются на разные интеллектуальные сферы (слова, числа, категории), подразумевают поиск, перебор вариантов. В первой задаче требуется найти способ получения следующего элемента последовательности по предыдущим (или формулу для n-ного элемента последовательности). Во второй и третьей задаче требуется найти способ получения указанного в скобках слова или числа и применить этот способ к незаполненному примеру. При этом в третьей задаче способ достаточно очевиден, однако само его применение требует некоторого перебора. В пропущенном слове на первой позиции может оказаться как «Г», так и «Ж», а на последней — как «А», так и «К».

Четвертая и пятая задачи также требуют определенного перебора. В четвертой задаче необходимо перебирать возможные перестановки букв в предложенных словах, пытаясь получить фамилии поэтов (либо перебирать фамилии поэтов и сравнивать их с предложенными наборами букв). Интересно, что если бы не указывалась принадлежность данных наборов букв к множеству фамилий поэтов (а например, просто говорилось «исключите лишнее»), задача решалась бы гораздо сложнее. В частности, подобрать правильное слово к набору, не являющемуся фамилией, в этой задаче сложнее, чем к остальным наборам. Пятая задача тоже решается перебором: можно перебирать недостающие буквы (или слова целиком) в поисках такой комбинации, которая будет удовлетворять условию. То, что человек может «увидеть» ответ без сознательного перебора, говорит

не о том, что человек находит ответ в результате какого-то мистического озарения, а скорее, о том, что какой-то процесс поиска идет не на уровне сознания.

Итак, в большинстве своем тесты предполагают не применение явной вычислительной схемы, а поиск решения в результате перебора вариантов. Действительно, если бы в первой задаче было сказано: продолжите последовательность, прибавляя к последнему ее элементу умноженный на два предыдущий элемент, — вряд ли мы бы подумали, что это задача на мышление (в лучшем случае — на знание математических операций). Или, скажем, если бы в четвертой задаче слова были представлены не в зашифрованном виде, то эту «задачу» мы посчитали бы тестом на знание фамилий поэтов. Наоборот, если бы были зашифрованы, например, латинские названия растений, такая «задача на мышление» нас бы возмутила, поскольку была бы задачей не столько на мышление, сколько на обладание специфическим знанием.

Таким образом, многие задачи в тестах на интеллект оказываются типичными примерами NP-полных задач, хотя их размерности являются достаточно малыми. Интересно, что даже в этих случаях человек не выполняет исчерпывающего поиска ответа. Так, в четвертой задаче вы не перебираете все варианты перестановок букв. Думая, например, над расшифровкой «Воскарен», можно предположить, что это фамилия русского поэта, заканчивающаяся на «ов», далее посмотреть, какая из оставшихся букв может быть первой в фамилии и повспоминать поэтов, чьи фамилии начинаются на эту букву. То, что человек не находит мгновенно ответы для подобных задач и применяет подобные приемы для сокращения перебора, еще раз подтверждает (хотя и не доказывает однозначно) отсутствие у человека мистических «невычислимых» способностей.

Хотя интеллект человека тестируется на задачах с малой размерностью, при разработке компьютерных программ возникает большой соблазн реализовать упрощенные алгоритмы, выполняющие исчерпывающий перебор вариантов. Из-за этого гораздо продуктивнее обращаться к задачам, которые не допускают полного перебора в силу большой размерности.

Но почему вообще у человека появилась способность к решению задач, вовлекающих поиск? Есть ли сходные задачи, возникающие перед предками человека или животными в естественных условиях?

МЫСЛЯТ ЛИ ЖИВОТНЫЕ?

На протяжении всей истории человечества людям регулярно приходилось взаимодействовать с животными: защита от хищников и охота, животноводство и одомашнивание вообще. Во всех этих случаях была потребность изучать повадки животных. Весьма давно было замечено, что повадки животных разных видов различаются, и в их поведении часто проявляются одинаковые для животных данного вида стереотипные действия. При этом порой такие действия выполняются животными, даже если при текущих обстоятельствах они оказываются бессмысленными (например, движения «закапывания» на твердой поверхности), что очень напоминает поведение примитивных автоматических механизмов. Такое поведение в III веке до н. э. стали называть инстинктивным и противопоставлять его индивидуальным особенностям поведения.

При субъективной оценке способностей животных всегда существовали две крайности: либо животным приписываются все человеческие качества, либо их лишают какой бы то ни было способности к рассудочной деятельности. В последнем случае полагают, что поведение животных полностью описывается инстинктами. Однако это не полностью верно даже для насекомых, хотя у них инстинктивное поведение развито, пожалуй, наиболее сильно.

Врожденность инстинктов была обоснована благодаря обширным наблюдениям за поведением животных, выросших в неволе, в зоопарке. В частности, бобры, выросшие в изоляции от сородичей, успешно строили хатки без какого-либо обучения или тренировки, чего не делали другие животные. Способность животных приобретать новые навыки, выходящие за рамки инстинктивного поведения, была наиболее явно обнаружена при их дрессировке.

Однако действительно научное изучение психики животных, по сути, началось лишь в XIX веке с работ Чарльза Дарвина. Не столько сходство в строении тела, сколько сходство некоторых форм поведения (например, мимики при выражении эмоций) заставило Дарвина сделать вывод об общности происхождения человека и обезьян. Это наблюдение стало одним из стимулов к созданию теории эволюции, в которой возникновение человеческого интеллекта, как и прочих свойств ныне существующих организмов, должно было происходить постепенно. Кроме того, Дарвин разделил неинстинктивные формы поведения на способность к обучению, вернее, к установлению ассоциаций, и рассудочную деятельность. Общность мыслительных процессов человека и животных и непрерывность развития умственных способностей в процессе эволюции обосновывается в книге 1888 года «Ум животных», написанной коллегой Дарвина, Джоном Роменсом.

Одной из первых содержательных теорий, направленных на объективное описание психики животных и человека, стало учение Ивана Петровича Павлова о высшей нервной деятельности. В рамках этого учения производилась попытка объяснить все формы поведения животных (и многие формы поведения человека) единообразно через рефлексы, работающие по схеме «стимул — реакция». Инстинкты представлялись как разновидность безусловных, наследственных рефлексов, на основе которых в результате работы механизмов, аналогичных установлению ассоциаций, строились условные рефлексы.

Объективность понятия рефлекса была продемонстрирована на примере многих видов животных. Особенности формирования условных рефлексов поддавались количественным оценкам, которые обладали хорошей повторяемостью от особи к особи. Данное учение не только сыграло большую роль в развитии физиологии, но и имело определенное значение для области искусственного интеллекта. Хотя условные рефлексы, надстраиваясь над инстинктами, дают новый уровень организации психики, они еще не являются самим мышлением, и переход от одного к другому в учении о высшей нервной деятельности не описывается (хотя более сложные виды рефлексов в нем присутствуют). Теория Дарвина пред-

сказывала (в отличие от «теорий» неэволюционного происхождения человека) существование уровней организации психики животных, которые бы заполнили разрыв между уровнем условных рефлексов и мышлением человека, и толкала ученых на их поиск.

Одновременно с Павловым объективным изучением поведения животных занимался Эдвард Торндайк, который первым ввел такую количественную характеристику, как кривая научения — зависимость сформированности некоторого навыка от числа попыток его использования. Он полагал, что интеллектуальное поведение может проявиться только в тех случаях, когда у животного нет готового (по сути, рефлекторного) ответа для текущей ситуации. Подобные ситуации он назвал *проблемными* и вместо выработки у животных условных рефлексов использовал метод проблемных ящиков, в котором животному вместо пассивного восприятия какого-то нового стимула нужно было совершать активные заранее неизвестные действия (например, отодвинуть задвижку у ящика с пищей). Посмотрим на следующие строки из книги З. А. Зориной и И. И. Полетаевой «Элементарное мышление животных»:

«В книге „Интеллект животных” (1898) Торндайк утверждал, что решение задачи является интеллектуальным актом. Решение задачи появляется как результат активных действий индивида благодаря последовательному перебору различных манипуляций».

Итак, по Торндайку, интеллект — это способность к нахождению действий, позволяющих разрешить проблемную ситуацию, для которой готового ответа нет, в то время как инстинкт (или рефлекс) — это применение известной последовательности действий. Теория интеллекта как поиска в это время также развивалась Клапаредом, который, однако, отмечал, что поиск не может быть чисто случайным. Однако в опытах Торндайка с проблемными ящиками животные были помещены в такие условия, когда они могли найти решение только *методом проб и ошибок*, поскольку решение никак не вытекало из структуры проблемной ситуации (например, дверца ящика отворялась при нажатии некоторой кнопки, видимым образом не связанной с дверцей).

Этот недостаток эксперимента был устранен в опытах Вольфганга Кёлера, проводившихся преимущественно с обезьянами. На основе своих опытов Кёлер убедительно обосновал, что поведение высших животных не исчерпывается условными рефлексами.

Кёлер ставил перед обезьянами (шимпанзе) задачи различной степени сложности. Одной из самых простых была задача, в которой к плоду, лежащему за решеткой, была привязана веревка, свободный конец которой находился перед решеткой. Все шимпанзе справлялись с этой задачей, сразу же хватая за веревку и подтаскивая к решетке плод. Легкость решения этой задачи обезьянами (в отличие, например, от собак, которые самостоятельно не могут «догадаться» потянуть за веревку) может быть связана с тем, что в естественных условиях притягивание плода, висящего на ветке, является инстинктивным.

Однако обезьянами уже не столь легко решается задача, в которой свободный конец веревки лежит за решеткой и до него невозможно дотянуться, а требуется приблизить его к себе с помощью лежащей неподалеку палки. Тем не менее это также выполнялось некоторыми обезьянами. Задача еще больше усложнялась, когда нужно было предварительно палку «изготовить», например составить длинную палку из двух недостаточно длинных трубочек тростника или отломать тонкую доску от большого ящика. Кёлером ставились и многие другие задачи. Сложные задачи могли решаться только одной, самой умной обезьяной и далеко не сразу, что свидетельствовало против того, что решение подобных задач осуществляется инстинктивно.

Кроме того, сам характер поведения животных в процессе решения не имеет ничего общего с жесткой схемой реализации инстинкта. Если животное сталкивается с ситуацией, для которой готов рефлекторный ответ, поведение животного выглядит удивительно целеустремленным и выверенным. Когда же такого ответа нет, животное приходит в сильное возбуждение, начинает совершать множество лишних, случайных действий: бегать вдоль решетки, пытаться просунуть в нее лапы и голову, а если плод подвешен высоко — прыгать к нему с разных точек.

Выдающийся отечественный ученый Лев Семенович Выготский, также работавший в этой области в первой половине XX века, отмечал, что подобное поведение свойственно многим животным. Даже муравей, идущий по феромонному следу, начнет беспорядочно бегать в разные стороны, если встретит на своем пути неожиданное препятствие. Сходным образом ведет себя и курица, когда отверстие в ограде, через которое она обычно подходила к кормушке, оказывается закрытым. Курица приходит в сильное возбуждение и начинает метаться вдоль ограды, совершая беспорядочные попытки найти подходящее отверстие. И даже более высокоорганизованные животные, такие как собаки, вполне могут проявлять аналогичное поведение, когда видят кусок мяса, являющийся непосредственно недостижимым.

Со стороны такое поведение может показаться абсолютно бессмысленным. Однако именно оно является прообразом мышления. Когда готового ответа нет, ничего не остается, как искать его, совершая случайные пробы, которые в конечном итоге могут помочь найти обходной путь, преодолеть препятствие. Стоит отметить, что и инстинкты обычно содержат поисковую фазу, в ходе которой животное ищет ключевой раздражитель или ситуацию, допускающую реализацию инстинктивной программы. Большинство животных совершает такой перебор в физическом пространстве, но у обезьян после серии неудачных проб перепроизводство движения прекращается, они садятся и замирают, фиксируя цель глазами. После этого они через некоторое время могут спокойно подняться и, выполнив правильную последовательность действий без новых проб, достать желанный плод. Перебор действий во внешнем пространстве переходит в какой-то внутренний, мысленный поиск, при котором происходит перебор цепочек возможных действий для разрешения проблемной ситуации. Это весьма примечательный факт, позволяющий подступиться к проблеме мышления.

Но как происходит этот поиск? Почему при решении задач обезьянами наблюдаются следующие особенности:

• если палка или другой инструмент лежит вне поля зрения обезьяны (но в той же комнате), догадаться применить ее обезьяне заметно труднее;

• задачи, при решении которых одним из правильных действий является отодвигание плода от себя, решаются гораздо труднее, чем задачи с придвиганием плода;

• даже небольшое дополнительное число промежуточных действий, необходимых для решения задачи, может сделать ее неразрешимой для обезьяны;

• обезьяна гораздо лучше справляется с задачами, если они предъявляются ей в порядке возрастания сложности.

Эти факты могут показаться достаточно тривиальными. Но как их объяснить содержательно, какие выводы об устройстве мыслительных процессов можно сделать? Мы эти вопросы пока отложим. Вместо этого посмотрим, применимо ли к человеку представление мышления как перебора цепочек возможных действий.

МЫСЛЬ И ДЕЙСТВИЕ

Обезьяна перебирает те действия, которые для нее являются известными, например притягивание ветки или использование палки. Если какое-то действие неизвестно (непривычно), например составление длинной палки из двух палок, одна из которых полая, то обезьяне к нему прибегнуть значительно сложнее. Животное берет обе палки в руки и держит за место соединения, пытаясь так получить длинную палку. Сама эта попытка, возникающая, когда есть несколько палок недостаточной длины, чтобы достать плод, свидетельствует о том, что обезьяна не просто реагирует каким-то двигательным ответом на текущую ситуацию, но пытается составить план действий для достижения цели. Однако догадаться вставить одну палку в другую ей очень сложно. Обычно это происходит случайно во время игры. Но когда операция соединения палок становится привычной, она легко применяется при решении задач. Таким образом, обезьяна в процессе мышления перебирает цепочки известных действий (инстинктивных или выученных — заимствованных при помощи подражания, случайно обнаруженных в игре и т. д.).

Еще в XIX веке И. В. Сеченов писал, что мышление — это «свернутое» (или задержанное) движение, действие. Не значит ли это, что увеличение многообразия доступных действий является необходимой основой для расширения возможностей мышления? Для нас действие — это не столько движение или мышечный акт сам по себе, сколько направленное взаимодействие с некоторым объектом. Действительно, «разумность» животных мы часто интуитивно оцениваем по их способности манипулировать предметами, которая является предпосылкой орудийной деятельности.

Возникновение орудийной деятельности сыграло существенную роль в развитии мышления. До ее появления глубокий перебор вариантов у животных был востребован разве что при поиске и планировании пути. А для этой частной задачи эволюцией вполне могли быть выработаны и частные решения, мало пригодные для других задач. Однако для многих животных оценка их уровня интеллектуального развития осуществлялась с использованием тестов на поиск пути в лабиринте. Под впечатлением от этих опытов была даже сформулирована *лабиринтная гипотеза мышления*.

Казалось бы, что общего между блужданиями крысы в лабиринте и человеческим мышлением, например при поиске доказательства математической теоремы? Но удивительная общность, связанная с универсальностью проблемы поиска, имеется. По сути, лабиринтная гипотеза отождествляет мышление с поиском, просто пространство поиска (не обязательно являющееся физическим пространством) уподобляется лабиринту. В этом лабиринте входом является условие задачи, выходом — ее решение, а пути в лабиринте определяются доступными действиями. Таким образом, задачи из IQ-тестов, как и задачи Кёлера для обезьян, требуют поиска или перебора последовательностей действий или манипуляций, что вполне соответствует утверждению Торндайка и лабиринтной гипотезе мышления. Но в чем же тогда отличие мышления обезьяны и человека?..

У ряда животных есть зачатки орудийной деятельности, но раньше считалось, что ни одно животное не занимается преднамеренным изготовлением новых орудий труда (на самом деле, и здесь бывают исключения). Как это ни банально

звучит, но именно труд сделал из обезьяны человека. При этом начало технического прогресса и развития цивилизации стало возможным не столько вследствие появления продуктов труда, сколько вследствие развития мышления, необходимого для осуществления этой деятельности. Но как можно проверить, будет ли на развитие мышления оказывать влияние расширение множества доступных действий, которое должно происходить за счет изготовления и использования орудий труда?

Еще одной наукой, получившей активное развитие в начале XX века, стала этнология. И хотя первобытных племен на Земле найти уже нельзя, все еще остаются относительно примитивные общества, которых мало коснулся технический прогресс. В жизни людей в этих обществах большую роль играет самый разнообразный ручной труд. Но как разобраться в особенностях мышления этих людей, если даже об устройстве собственного мышления что-либо сказать трудно?

Американский этнолог Ф. Кашинг тесно общался с индейским племенем зуньи. Пытаясь проникнуть в их жизнь и способ мышления, он, в частности, на протяжении длительного времени выполнял руками все те же операции и в тех же условиях, что и индейцы. По словам Кашинга, он вернул свои руки к тому первобытному состоянию, «когда они были так связаны с интеллектом, что действительно составляли его часть». И даже статья, в которой он в 1892 году опубликовал результаты своих исследований, была им озаглавлена «Ручные понятия».

Как указывает академик В. В. Иванов, опыт Кашинга повторил Эйзенштейн, испытав при этом, что «двигательный акт есть одновременно акт мышления, а мысль — одновременно — пространственное действие». Сходные ощущения, но в более слабой форме, могут ощутить люди, усиленно осваивающие какой-либо вид спорта или некоторых танцев, требующий сложных скоординированных действий в ответ на сложившуюся ситуацию. Некий набор «ручных понятий» можно также сформировать путем длительного решения механических головоломок. Тесную связь мышления и действия показывает также то, что в результате тренировки

двигательных функций у маленьких детей повышается и уровень интеллекта.

Таким образом, феномен «ручных понятий» показывает происхождение мышления из действия, причем возникновение труда существенно расширяет набор действий и обогащает мышление. Когда же мышление начинает оперировать символами и операциями над ними вместо конкретных действий, его возможности безгранично расширяются, поскольку оно уже не сковывается тесными рамками движений тела в физическом пространстве.

Использование языка вполне естественным (но вовсе не тривиальным) образом расширяет мышление как «свернутое» действие. Языковое мышление, видимо, имеет ту же природу, что и «ручное» мышление. Исходно для маленького ребенка речь — это такое же физическое действие, как, например, для обезьяны — подтягивание ветки с плодом, поскольку произнесение какого-то слова (например, «няня») позволяет приблизить желаемый предмет или удовлетворить какие-то потребности (причем меняющиеся в зависимости от ситуации). Иногда такое использование языка, направленного на удовлетворение органических потребностей, называется *аутической речью*.

Позднее (в возрасте 3–5 лет) ребенок может часто говорить, не обращаясь к кому-либо конкретно. Такую речь Ж. Пиаже назвал *эгоцентрической*, полагая, что она является промежуточной ступенью на пути к социализированной речи. Однако позднее Пиаже согласился с критическим анализом Выготского, в котором показывалось, что эгоцентрическая речь несет в себе функцию мышления и является ступенью на пути не к социализированной, а напротив, внутренней речи, которая часто нами воспринимается как собственно мышление. Но даже при внутренней речи мозг продолжает посылать слабые сигналы на органы речи, что достоверно фиксируется приборами.

Сходным образом, видимо, формируются и «ручные понятия». Сначала человек просто производит конкретные действия, как-то влияющие на окружение. Затем эти действия начинают осуществляться беспредметно в воображаемой ситуации (как в случае эгоцентрической речи), и постепенно

они «свертываются» (как и в случае перехода ко внутренней речи), формируя специфический «язык», на котором человек может мыслить.

Между физическим действием и языком, да и мышлением в целом, есть глубокая связь. Однако, как отмечалось, мышление тождественно не просто свернутому действию, но поиску оптимальных цепочек действий, которые бы позволяли достичь некоторого желаемого результата в проблемной ситуации. В случае с физическими действиями такая постановка кажется достаточно естественной, но является ли она правдоподобной применительно к символьному мышлению? В математике, как мы видели, решение задачи в общем виде описывается в форме алгоритма, превращающегося в нужную цепочку операций над символами в зависимости от конкретных условий задачи («линейная» последовательность действий является частным случаем алгоритма). Чтобы решить задачу, нужно найти подходящий алгоритм, но соответствует ли такое представление процессу решения задачи человеком? Кроме того, множество цепочек действий может расти экспоненциально с увеличением их длины, а если требуется найти не просто конечную цепочку действий, но алгоритм, то задача может оказаться и вовсе неразрешимой. Лабиринтная гипотеза выглядит вполне правдоподобной, но она не говорит о механизмах поиска, которые, видимо, могут быть весьма нетривиальными. Тем не менее проблема поиска — это та ниточка, за которую стал распутываться клубок тайн вокруг мышления.

РАССУЖДЕНИЯ О МЕТОДЕ

Формально-аксиоматический подход позволил представить математические задачи и их решения в символьной форме, однако сам процесс вывода доказательств (поиска алгоритма), осуществляемый математиками, оставался неформализованным. Из-за отсутствия окончательной формализации ученым по-прежнему приходилось пользоваться нестрогими методами, применение которых являлось определенного рода искусством.

Исследование подобных методов проводил еще Декарт в своей книге «Правила для руководства ума». Впоследствии, как уже отмечалось, идея о выработке «правил для руководства ума» при поиске решения благодаря Лейбницу трансформировалась в идею построения машин, выполняющих решение автоматически, т. е. в идею полной формализации этих правил. Поскольку полное и окончательное решение данной проблемы оказалось непосредственно недостижимым, некоторые ученые вернулись к рассуждениям о методе в стиле Декарта. В этом направлении наибольший интерес представляет детальный анализ, проведенный математиком и педагогом Д. Пойа на примере решения задач школьниками. Ведь проблема метода математического доказательства возникает не только у профессиональных математиков. В равной мере она встает и перед школьниками, вернее, перед учителями, занимающимися их обучением.

Пойа подчеркивал, что владение неким предметом подразумевает обладание не только знаниями, но и в большей степени умением их применять. Отсутствие соответствующих навыков приводит к полной неспособности ученика хоть как-то приступить к решению задачи (даже в том случае, когда готовое доказательство вполне может быть им понято) либо, в лучшем случае, к использованию метода «научного тыка», т. е. к случайному перебору разных комбинаций возможных действий над исходными данными. Вывод вполне прост: само доказательство и процесс его поиска — совершенно разные вещи.

Убедимся в этом на примере следующей задачи, приведенной Пойа в его книге «Математическое открытие»: найти сумму квадратов N первых натуральных чисел $1^2 + 2^2 + \ldots + N^2$. Ответ у этой задачи следующий

$$\sum_{n=1}^{N} n^2 = \frac{2N^3 + 3N^2 + N}{6}.$$

Интересно уже то, что доказать правильность этого ответа проще, хотя и ненамного, чем решить исходную задачу. Как будто мы знаем общее направление к выходу из лабиринта, но вовсе не знаем пути к нему. Если вы не помните вывода

этого ответа, можете попытаться найти его самостоятельно, прежде чем рассмотреть следующее решение:

$$\sum_{n=1}^{N} n^2 = \frac{1}{3}\sum_{n=1}^{N} 3n^2 = \frac{1}{3}\sum_{n=1}^{N} (-n^3 + n^3 + 3n^2) =$$

$$= \frac{1}{3}\sum_{n=1}^{N} (-n^3 + n^3 + 3n^2 + 3n + 1 - 3n - 1) =$$

$$= \frac{1}{3}\sum_{n=1}^{N} ((n+1)^3 - n^3 - 3n - 1) =$$

$$= \frac{1}{3}\sum_{n=1}^{N} ((n+1)^3 - n^3) - \frac{1}{3}\sum_{n=1}^{N} (3n + 1) =$$

$$= \frac{(N+1)^3 - 1}{3} - \frac{3N(N+1) + 2N}{6} = \frac{2N^3 + 3N^2 + N}{6}.$$

Истинность всех преобразований очевидна (кроме, возможно, преобразования, после которого пропадают знаки суммирования, но в его корректности тоже несложно убедиться). Практически любой может проверить правильность данного доказательства, что является бесспорной заслугой современной символьной математики. Задача решена, ... но это решение оставляет определенную неудовлетворенность, поскольку от него хочется чего-то большего, чем просто подтверждения приведенной выше формулы. Это решение приведено как данность, и остается совершенно не ясно, из каких соображений выбирались именно такие преобразования и как до них можно додуматься самостоятельно.

Не получено же оно случайным перебором всех возможных преобразований исходного выражения с поиском такой цепочки преобразований, которая позволит получить ответ (хотя кто-то достаточно терпеливый может решить эту задачу и таким путем)? Данное решение уже описывает путь от входа в лабиринт к выходу из него, но не сообщает нам, как этот путь был найден. Значит, навыки решения задач представляют собой совокупность приемов или методов, позволяющих направить поиск решения.

Пойа проводил исследования методов решения задач, которые он назвал «эвристическими». Это название является

производным от слова «эврика», в переводе с греческого значащего «нашел», или «открыл». Именно с таким возгласом по легенде Архимед выскочил из ванны, придумав, как определить материал короны царя Сиракуз. Термин «эвристический» может трактоваться как «способствующий открытию». Хотя это значение данный термин продолжает сохранять, сейчас он приобрел дополнительный оттенок — «нестрогий, придуманный на основе интуитивных соображений». Под эвристикой в области ИИ понимается прием или метод, облегчающий поиск решения задачи (именно в этом смысле он и будет использоваться далее).

В своей книге Пойа подчеркивал единство методов решения школьных задач и научного поиска. Цитируя Г. Спенсера, он писал: «Что значит преподавать? Это значит систематически побуждать учащихся к собственным открытиям». Кроме того, Пойа тоже проводил аналогию между поиском решения произвольной задачи и поиском пути в ограниченно известной местности как наиболее типичной задачей, возникающей перед животными и первобытными людьми, благодаря чему мы затруднения при решении любых задач воспринимаем, *как препятствия на пути к достижению цели.*

Неужели все задачи от поиска пути на местности до свершения научных открытий имеют что-то общее, и для их решения может использоваться один и тот же интеллект? Если действительно модель мышления как направленного поиска может описать решения самых разнообразных задач, то эта модель должна отражать самую базовую сущность мышления.

Но какие все же приемы помогают найти решение задачи или, тем более, сделать научное открытие? Рассмотрим для примера следующую простую задачу: дана окружность; с помощью циркуля и линейки необходимо найти положение ее центра.

Прежде чем приступать к решению задачи, необходимо установить набор допустимых действий. Мы можем:

• провести через заданные две точки прямую;

• провести окружность заданного радиуса с центром в заданной точке;

• установить раствор циркуля по двум точкам;

• определить точку пересечения двух линий (прямых или окружностей).

Есть еще и «дополнительные» разрешенные действия:

• случайным образом провести прямую через одну точку;

• случайным образом выбрать одну из точек на заданной геометрической фигуре (например, окружности);

• случайным образом уменьшить или увеличить раствор циркуля.

Наличие последних трех действий крайне важно, хотя о них часто в явном виде не упоминается. С другой стороны, некоторые относительно естественные действия запрещены. Например, нельзя «на глазок» провести касательную к окружности или, понемногу увеличивая раствор циркуля, найти наиболее удаленные точки на окружности.

Как можно решить задачу с поиском центра окружности? Для начала у нас немного альтернатив: можно выбрать случайно точку на окружности и провести через нее случайную прямую или окружность со случайно заданным радиусом. После этого мы можем получить новые точки на исходной окружности и использовать их для следующих построений. Число возможных построений будет расти чрезвычайно быстро.

Человеку, не имеющему опыта решения таких задач, будет крайне сложно найти правильную последовательность действий, несмотря на то, что эта последовательность весьма коротка. Если же задачу усложнить, то ждать ее решения и вовсе бесперспективно. Чтобы человек смог решать такие задачи, у него сначала должны сформироваться навыки решения на постепенно усложняющихся примерах (причем желательно, чтобы процесс решения сопровождался комментариями учителя). Но что именно дают эти навыки?

Решение задачи с нахождением центра окружности является простым, если владеть некоторыми дополнительными приемами, такими как проведение перпендикуляра через середину отрезка, и знать, что перпендикуляр, проходящий через середину хорды окружности, проходит и через ее центр (этот факт действительно нужно знать: он легко получается в аналитической геометрии, но доказать подобные факты в

рамках построений с помощью циркуля и линейки проблематично). Описание решения с использованием дополнительных действий выглядит очень компактным: «центр окружности может быть найден как точка пересечения перпендикуляров, проходящих через середины двух произвольных хорд». Представьте, насколько длиннее было бы описание решения в терминах элементарных действий с циркулем и линейкой. По сути, деление отрезка на две равные части и построение к нему перпендикуляра (да и сами понятия деления отрезков и перпендикуляров) являются частными эвристиками, существенно облегчающими поиск решения исходной задачи. Операция построения перпендикуляра к отрезку по своему значению сходна с операцией составления длинной палки из двух при решении задач на доставание плода обезьяной. Разница лишь в том, что первая — символьная.

Перпендикуляр через середину отрезка строится всего в четыре действия: 1) устанавливается раствор циркуля по концам отрезка; 2) и 3) проводятся окружности выбранного радиуса с центрами в концах отрезков; 4) через две точки их пересечения проводится прямая.

Существуют тысячи других цепочек из четырех действий (даже если ограничиться элементарными действиями с циркулем и линейкой), которые для большинства задач оказываются малополезными, однако в исключительных случаях могут быть применимы. Именно в этом суть эвристик: они обычно помогают находить решения, хотя при этом не дают полной гарантии. Существуют задачи, в которых требуются сложные вспомогательные построения; если всецело полагаться только на освоенные приемы, то решения таких задач найти не удастся.

Задачу о построении перпендикуляра многие школьники способны решить самостоятельно за приемлемое время — интуитивно или сознательно выполняя перебор цепочек доступных действий. После этого задача о центре окружности также становится решаемой. А вот без дополнительных эвристик она практически не решаема, поскольку правильная цепочка элементарных действий затеряна в миллиардах других потенциальных построений (кто-то, возможно, сможет решить и эту задачу, не зная о существовании перпендикуляров, но

задействовав какие-то иные эвристики). Здесь важно знать даже не сам способ построения перпендикуляра, а то, что его в принципе можно построить и использовать в данной задаче. Ведь поиск цепочки действий для выполнения одного заведомо осуществимого вспомогательного построения значительно проще, чем перебор всех потенциальных построений, многие из которых могут быть и неосуществимы, что установить крайне трудно как математически, так и на основе здравого смысла. Вряд ли кто-то, не знающий заранее ответа, может с уверенностью сказать, какие из следующих построений возможны, а какие — нет: 1) построение окружности, проходящей через все вершины произвольного треугольника; 2) построение квадрата, равного по площади данному кругу; 3) разделение данного угла на три равные части. Поиск некоторых построений (например, квадратуры круга) шел на протяжении многих веков, пока их неосуществимость не была доказана алгебраически.

Новичок не может сразу решать сложные задачи, сначала он должен освоить (или постепенно самостоятельно изобрести) массу эвристик, специфичных для данного класса задач. Подобные *предметно-зависимые* эвристики, однако, не устраивали математиков, искавших общие методы решения задач.

Такой общностью обладают, например, метод разделения задачи на подзадачи с независимым их решением, метод поиска решения для частного случая с последующим обобщением или метод аналогий. Эти методы, однако, не удается сформулировать в виде легко применимых четких правил, и оказывается необходимым приобретать навыки их практического использования.

Исследования Пойа так и не привели к созданию нового направления в математике или педагогике, однако его рассуждения о методах решения задач, подкрепленные «лабиринтной гипотезой» в психологии и работами математиков в области теории алгоритмов сыграли свою роль при формировании первого общего подхода к проблеме искусственного интеллекта. Этот общий подход получил название «эвристическое программирование».

К моменту появления компьютеров разрозненные данные психологии, этнологии, зоопсихологии, педагогики и методологии науки свидетельствовали о том, что интеллект удобно представлять как способность к решению задач (проблемных ситуаций), а сам процесс решения (мышление) реализуется как поиск такой цепочки действий, которая начальную ситуацию (условие задачи) переводила бы в конечную ситуацию (ответ). Начиная с некоторого уровня развития живых организмов, перебор действий путем их фактического исполнения заменяется «виртуальным» перебором «свернутых» действий. В математике в это время в целях строгого описания правильных рассуждений было формализовано понятие алгоритма, развертывающегося в зависящую от исходных данных последовательность действий и описывающего решение задачи в общем виде. Компьютеры были созданы как устройства, автоматически выполняющие алгоритмы, т. е. решающие задачи, представленные в символьной форме. Однако оставался ключевой вопрос: как находить сами алгоритмы или хотя бы линейные цепочки последовательных действий? Из-за NP-полноты или даже неразрешимости общей проблемы поиска полный перебор всех возможных вариантов во многих случаях оказывается неосуществимым.

Согласно Пойя, человеку удается в какой-то мере преодолевать данную проблему, используя эвристики — приемы, облегчающие решение задач. Но можно ли эвристики запрограммировать? А если можно, в каком виде это следует делать? Попытка ученых ответить на эти вопросы привела в 1950–60-х годах к возникновению одной из первых парадигм искусственного интеллекта, получившей название «эвристическое программирование». В области ИИ с момента ее возникновения существовало много различных направлений исследований, но именно эвристическое программирование можно охарактеризовать как первую попытку создания действительно общей теории искусственного интеллекта.

В эвристическом программировании проблема искусственного интеллекта часто изучалась на примере интеллектуальных игр, таких как шашки, шахматы и т. д. С одной стороны, «вселенная» любой настольной игры задается маленьким набором очень простых законов, которые легко запрограммировать. С другой стороны, в процессе игры, к примеру в шахматы, безусловно, необходимо думать. Задачи доказательства математических теорем и задачи планирования также широко исследовались в эвристическом программировании, однако они требовали дополнительных предметных знаний (к примеру, объяснить человеку правила игры в шахматы гораздо проще, чем правила интегрирования) и были менее показательными и более сложными в реализации. Попробуем сначала посмотреть на проблему мышления на примере простой игры.

Приведенные ниже рассуждения могут показаться тривиальными в силу того, что хорошо известны и более сложные методы. Однако представьте себя на месте первооткрывателей, которым абсолютно ничего еще не было известно и неоткуда было позаимствовать готовые решения. Это поможет лучше понять предпосылки и пути развития области эвристического программирования и исследований искусственного интеллекта вообще. Ведь среди привычных готовых решений многое может быть случайным и не оптимальным, в связи с чем неплохо пересматривать старые результаты в более широком контексте современных представлений.

Пусть перед нами стоит задача научить компьютер играть в простейшую интеллектуальную игру — «крестики-нолики» на поле 3×3. Как бы вы стали решать такую задачу?

Первое желание состоит в том, чтобы в явном виде указать, какие ходы должен делать компьютер. И новички обычно идут именно по этому пути, поскольку он кажется самым простым и контролируемым. Пусть компьютер играет крестиками. Тогда можно явно задать, что первый ход делается в центр поля. Далее, однако, ответ компьютера должен зависеть от хода игрока.

Обозначим клетки доски в «шахматном» стиле: от A1 до C3. Тогда такой алгоритм игры (с использованием синтаксиса языка C++) может выглядеть примерно так

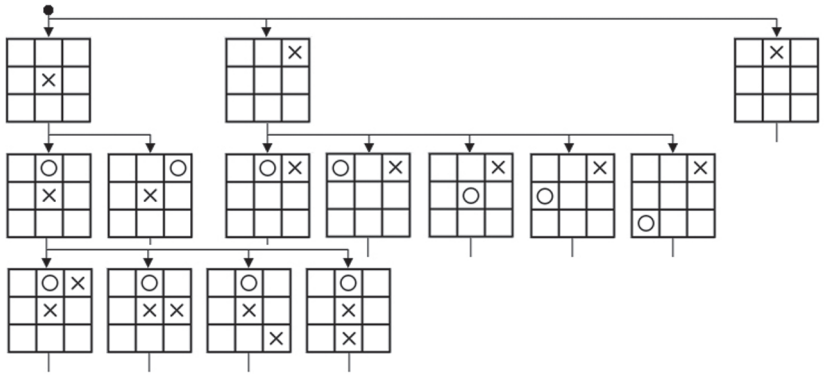
```
make_move ('X', "B2");
if (last_move ('O') == "A1") {
```

```
    make_move ('X', "C2");
    if (last_move ('O') == "A2") {
      make_move ('X', "A3");
      if (last_move ('O') == "C1") {
        make_move ('X', "B1");        // ... ничья
      } else {
        make_move ('X', "C1");        // победа
      }
    } else {
      make_move ('X', "A2");// победа
    }
  } else ...
```

Такой алгоритм сродни системе простейших рефлексов. Однако откуда берутся ходы, записанные в нашей программе? Чтобы убедиться в том, что некоторый наш ход приводит к выигрышу, нужно проверить все возможные ходы противника. Это можно представить в виде *дерева игры*, также называемого *деревом вариантов* для неигровых задач. На следующем рисунке показан пример фрагмента дерева игры для «крестиков-ноликов» на поле 3×3. Корень этого дерева представляет собой пустую доску, а выходящие из него ветви соответствуют возможным ходам первого игрока. Эти ветви ведут в новые состояния игрового мира, из каждого из которых также выходят ветви, соответствующие разрешенным ходам второго игрока, и так далее. В случае неигровых задач в корне дерева вариантов может находиться, например, решаемое уравнение, а ветви будут соответствовать допустимым операциям по преобразованию этого выражения.

По сути, такое дерево описывает *пространство состояний* — совокупность состояний, в которых может находиться некоторый, например игровой, мир с указанием возможных переходов между состояниями. Для некоторых задач это пространство может задаваться путем явного перечисления всех возможных состояний и переходов между ними, но чаще оно задается неявно — в форме правил (игры, интегрирования и т. д.). На основе правил может быть организована *порождающая процедура*, явно строящая дерево вариантов.

Мы пользуемся этой процедурой, чтобы построить дерево игры и найти на нем такие ответы, которые бы приводили



Пример фрагмента дерева игры

к победе или, если это невозможно, к ничьей. Их мы и заносим в нашу программу. А нельзя ли, чтобы компьютерная программа, играющая в «крестики-нолики», сама строила дерево вариантов на основе известной порождающей процедуры? Естественно, более элегантный (чем представленный выше) алгоритм и должен был бы сам строить дерево игры и выбирать наилучший ход, как это делаем мы.

Осуществить такое построение без хранения в памяти дополнительной копии игрового поля (а в более общем случае — модели мира) проблематично. Пусть field[3][3] — вспомогательный массив 3×3, значение в ячейке которого установлено в '–', если соответствующая клетка не занята, и в 'X' и 'O', если там «крестик» и «нолик» соответственно. Тогда порождающая процедура может быть записана в следующей форме:

```
char field[3][3];
void tree_search( char player )
{
  for( int i = 0; i < 3; i++ ) {
    for( int j = 0; j < 3; j++ ) {
      if( field[i][j] == '–') {
        field[i][j] = player;
        if( player == 'X')
          tree_search( 'O' );
        else
          tree_search( 'X' );
```

```
            field[i][j] = '-';
          }
        }
      }
    }
```

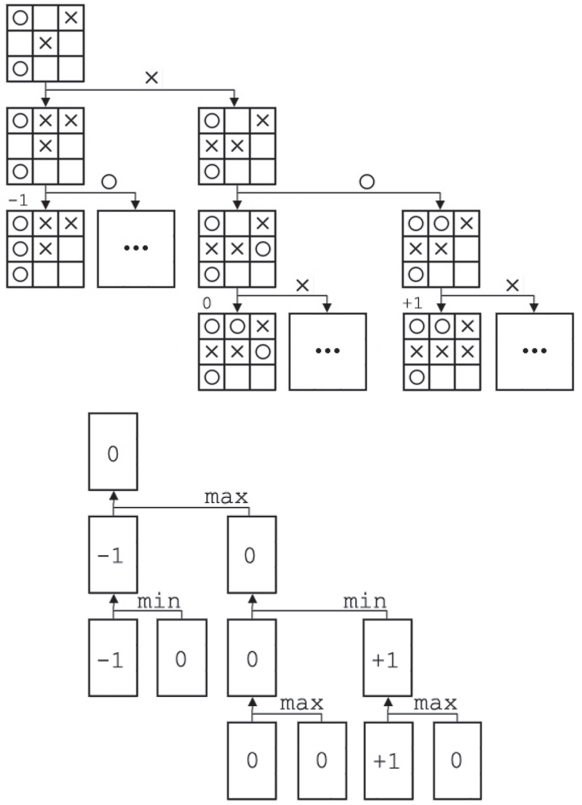
Если предварительно установить все значения массива field как пустые ('–'), а потом запустить эту процедуру tree_search('X'), то в процессе ее работы в массиве field поочередно окажутся все возможные комбинации расположения крестиков и ноликов.

Пусть имеется процедура check_game, которая оценивает ситуацию на доске, возвращая 'X' или 'O', если у соответствующего игрока поставлено 3 в ряд, либо '–', если такого нет. Для листьев нашего дерева мы можем установить их статус (чей-то выигрыш либо ничья).

Теперь рассмотрим узлы дерева, непосредственно предшествующие листьям. Если хотя бы один из листьев приводит к выигрышу игрока, ходу которого соответствует текущий узел, то данный узел также является выигрышным. Узел будет являться проигрышным только в том случае, если все листья, в которые он ведет, являются проигрышными. Таким способом можно пометить все узлы, предшествующие листьям, и продолжать распространять эту информацию вниз к корню.

Удобнее представить выигрыш компьютера как «+1», ничью как «0», а проигрыш компьютера как «–1». Тогда для определения статуса узла, соответствующего ходу компьютера, нужно будет брать максимум, а для хода противника — минимум по всем дочерним узлам. Такая процедура оценки статуса узлов, показанная на следующем рисунке, называется процедурой *минимакса*. Встроить процедуру минимакса в порождающую процедуру не представляет трудности.

Существует теорема о минимаксе, доказанная Джоном фон Нейманом (до эмиграции из Венгрии его имя было Янош Лайош Нейман), являющимся к тому же автором получившей распространение архитектуры компьютеров. Эта теорема говорит о существовании оптимальной стратегии для игр с нулевой суммой, таких как «крестики-нолики» или шахматы, в которых выигрыш одного игрока равен проигрышу другого игрока.



Пример применения процедуры минимакса

Необходимо отметить существование в теории игр более общего понятия — *равновесия Нэша*, соответствующего таким стратегиям игроков, от которых ни одному из них невыгодно отклоняться, если этого не делают и другие игроки.

Если порождающая процедура известна и с помощью процедуры минимакса можно определить оптимальный ход, в чем же тогда проблема? Как уже не один раз отмечалось выше, проблема — в NP-полноте. Иными словами, полное дерево вариантов может получиться чрезвычайно больших размеров: зачастую число узлов растет экспоненциально с глубиной дерева, поскольку каждый узел ведет к некоторому числу новых узлов, каждый из которых, в свою очередь, опять ведет к некоторому числу новых узлов. К примеру, размер дерева для шашек составляет 10^{40}, для шахмат — 10^{120},

а для го и совсем невообразимое число — 10^{400}. Забавно, что часто, чтобы подчеркнуть величину этих чисел, говорят об их существенном превосходстве количества атомов или элементарных частиц во Вселенной, не вполне правомерно сравнивая число объектов и число комбинаций состояний объектов. Однако это сравнение становится оправданным, если говорить о классических системах с массовой параллельностью: даже если каждый атом во Вселенной будет являться отдельным процессором, перебирающим 10^{100} комбинаций в секунду, работы всей Вселенной на протяжении всего времени, прошедшего с момента Большого взрыва, не хватит, чтобы сыграть идеальную партию в го.

Дерево вариантов, заданное в неявной форме через правила игры, является как бы невидимым для программы. Как крысе, чтобы обследовать реальный лабиринт, нужно потратить некоторое время на исследование каждой развилки, за которой продолжение пути не видно, так и программе нужно потратить некоторое время (вычислительные ресурсы), чтобы обследовать дерево вариантов и найти нужный путь к победе.

Несмотря на кажущуюся элементарность, игра в «крестики-нолики» даже на поле 3×3 для ребенка, не знакомого с ней, требует заметных умственных усилий. Эта игра требует перебора вариантов на глубину 4–5, что близко к сложности задачи о нахождении центра отрезка с помощью циркуля и линейки. Однако, «решив» один раз «крестики-нолики» на поле 3×3, мы будем знать это решение (путь в «лабиринте» вариантов), и игра потеряет для нас интерес. Для «взрослых» интеллектуальных игр такая ограниченность не свойственна, поскольку все пространство поиска в них не может быть обследовано в течение всей жизни человека и даже в течение всего времени существования цивилизации.

Что же можно сделать, если все дерево порождено быть не может? Из каких соображений выбирать какой-либо ход? В области эвристического программирования термин «эвристика» уточняется как прием, позволяющий в ходе перебора отсекать неперспективные ветви на дереве вариантов, что является заметно более конкретным определением, чем определение эвристики как приема, облегчающего решение задачи.

Одна эвристика уже неявно использовалась при составлении дерева игры для «крестиков-ноликов». На нем показаны все начальные игровые позиции, не переводящиеся друг в друга вращением и зеркальным отражением. Иными словами, использована эвристика симметрии. Сколько существует вариантов игры, если не учитывать симметрии? Их количество — не более чем факториал числа 9. Благодаря использованию симметрий это число уменьшается на порядок. Однако для больших полей относительный выигрыш оказывается значительно меньше, так как уже после выполнения нескольких ходов симметричные продолжения перестают существовать. Симметрия является очень общей эвристикой, однако не слишком мощной, поскольку симметрия игровой ситуации (даже если она имеется) очень быстро нарушается.

Можно представить и более сложную (и более частную) эвристику: выгоднее тот ход, который потенциально может использоваться в большем числе построений трех в ряд. По этой эвристике исходно для «крестиков» самым выгодным является ход в центр, поскольку он может участвовать в четырех разных построениях. Затем идут ходы в углы — для них число построений равно трем. Для ходов у стенок это число всего лишь два. Понятно, что этой эвристики недостаточно, поскольку она не учитывает возможности противника (например, завершить свою комбинацию).

Вполне естественным является желание оценить выгодность хода количественно. На основе такой количественной оценки можно было бы отсекать часть ветвей в каждом узле дерева вариантов. И, действительно, в эвристическом программировании задание эвристик в форме *(статической) оценивающей функции* является наиболее распространенным.

НАПРАВЛЕННЫЙ ПОИСК

Что может представлять собой оценивающая функция? Наиболее типичным примером в шахматах является разница в суммарной силе фигур игрока и противника: присвоив каж-

дой фигуре некоторый вес, можно оценить качество текущей позиции. Эта оценка вполне эвристична: она действительно облегчает выбор хода, но при этом вовсе не гарантирует того, что ход, приводящий к ситуации с наибольшим перевесом по фигурам, будет лучшим. Данная оценка является нестрогой, ее нельзя получить аналитически (дедуктивно). Если бы оценивающая функция позволяла непосредственно выбирать гарантированно оптимальный ход, это означало бы то, что задача не является NP-полной и для нее существует явная выигрышная стратегия без перебора вариантов.

Постоянный выбор хода, максимизирующего оценивающую функцию, не гарантирует того, что не возникнет ситуации, в которой все продолжения будут вести к ее уменьшению. Как будто мы поднимались все время в гору, но вдруг натолкнулись на обрыв. Такое свойство оценивающей функции называют немонотонностью, при которой появляются локальные максимумы.

Выбор хода может производиться на основе только оценивающей функции. Алгоритмы, осуществляющие такой выбор, называются *жадными*. Смысл данного названия вполне очевиден, если представить себе игру программы, всегда выбирающей ход, приводящий к съеданию самой сильной фигуры противника (если же ни один ход не позволяет ничего съесть, то может выбираться первый попавшийся ход), невзирая ни на какие другие факторы. Такое поведение, если оно проявляется человеком, выглядит донельзя жадным и недальновидным. Жадный алгоритм не видит дальше одного шага, как человек, пытающийся подняться непроглядной ночью на самую высокую горную вершину, но остановившийся наверху маленького холмика и не желающий с него спускаться, поскольку любое дальнейшее движение ведет только вниз, а направления на следующее возвышение не видно.

Конечно, можно усложнить оценивающую функцию каким-то образом учтя и позиционное преимущество, что сделает рельеф этой функции более монотонным и улучшит игру жадного алгоритма. На практике жадные алгоритмы действительно могут быть достаточными при решении некоторых задач, особенно, если удается подобрать хорошую оценивающую функцию: они позволяют выбирать ход (или,

говоря более обще, действие) гораздо эффективнее, чем при случайном выборе, при этом оставаясь не слишком ресурсоемкими.

Для борьбы с немонотонностью статической оценивающей функции и выбора более хорошего (по сравнению с жадным алгоритмом) хода нужно смотреть на оценку качества не ближайших узлов, достижимых за один ход, но более далеких узлов. Другими словами, можно сказать, что просмотр дерева вариантов от текущей позиции на некоторую глубину может позволить улучшить значения, даваемые оценивающей функцией. В связи с этим вводится такое понятие, как *процедура формирования рабочих оценок*. Под рабочей оценкой понимается качество некоторого узла на дереве вариантов, полученное на основе статических оценок качества его потомков. Если мы из узла, соответствующего текущей игровой ситуации, просмотрим не только его непосредственные дочерние узлы, но и их потомков, то сможем уточнить оценку качества дочерних узлов, например с помощью процедуры минимакса. Если раньше мы рассматривали процедуру минимакса, начинавшую распространение точных значений качества с конечных позиций (листьев), определяемых правилами игры, то теперь расширяем эту процедуру на случай промежуточных позиций и нестрогих оценок. Использование процедуры минимакса совместно с оценивающей функцией вносит дополнительную эвристичность, поскольку, помогая при выборе хода, не является строго обоснованным.

Действительно, суть минимаксной оценки в том, что использующая ее компьютерная программа ожидает от противника наиболее сильного (причем по оценке самой программы) хода. Минимаксная оценка, на первый взгляд, может показаться единственно правильной. Однако нужно понимать, что данная оценка не является логически обоснованной и содержит элемент эвристичности. В нее заложено предположение, согласно которому противник «думает» так же, как и компьютерная программа. Это не что иное, как модель противника. При отсутствии дополнительной информации о противнике такая модель может быть принята по умолчанию. Однако она не учитывает ни стиль игры противника, ни его силу. Это наиболее ярко проявляется при игре

с форой против как более сильного, так и более слабого противника. В утрированном случае компьютерная программа, давшая фору противнику и способная осуществить полный перебор, должна была бы сразу сдаться, если бы она руководствовалась минимаксной оценкой. Сейчас начинает активно развиваться область *«эксплуатирования противника»* (opponent exploiting), т. е. использования его уязвимостей. В рамках этой области исследуется вопрос, что нужно делать игроку, когда его противник не придерживается оптимальной (точнее, равновесной) стратегии. В частности, показано, что если сам игрок в этой ситуации будет придерживаться равновесной стратегии, то он не получит наибольшего потенциального выигрыша. Однако такж установлено, что отказ от равновесной стратегии в пользу попытки «эксплойта» не может гарантировать успех, так как сам игрок может стать объектом встречного «эксплойта». Для безопасного отказа от равновесной стратегии необходимо иметь очень надежную модель противника, что выходит далеко за рамки теории игр и эвристического программирования.

Несмотря на определенное упрощение, минимаксная оценка весьма удобна. Однако, чтобы применить процедуру минимакса для формирования рабочих оценок, необходимо главное: определить, какие из узлов дерева вариантов нужно посетить, чтобы на их основе улучшить оценку качества ходов, возможных из текущей ситуации.

В простейшем случае перебор осуществляется на фиксированную глубину, одинаковую для всех потомков. В зависимости от глубины поиска он может варьироваться от жадных алгоритмов (в которых процедура формирования рабочих оценок совпадает со статической оценивающей функцией) до полного перебора (в которых оценивающая функция заменяется точным значением качества конечных позиций). Как отмечалось, эффективность работы эвристической программы можно улучшать за счет усложнения оценивающей функции. Естественно, увеличение глубины перебора также делает рабочие оценки более монотонными, чем значения статической оценивающей функции.

Однако процедура поиска может быть улучшена и без увеличения числа просматриваемых потомков. Ведь смысл

оценивающей функции как эвристики не в непосредственном выборе действия, а в отсечении ветвей на дереве вариантов в процессе поиска. Следующими по сложности процедурами поиска являются процедуры *направленного сокращения.* Процедуры формирования рабочих оценок этого типа просматривают не всех потомков данного узла, а только некоторых из них в соответствии с их статическими оценками. За счет того, что неперспективные ветви отсекаются, перспективные ветви могут быть исследованы глубже при тех же вычислительных затратах. *Процедура n-наилучшего направленного сокращения* для каждого узла просматривает *n* потомков с максимальными (или минимальными при ходе соперника) значениями статической оценивающей функции.

Для существенного увеличения глубины поиска его, однако, необходимо сильно сужать. Если, например, в игре го на каждом ходе есть выбор между, скажем, ста вариантами постановки камня, то при просмотре лишь десяти наиболее перспективных ветвей в каждом узле глубину поиска можно увеличить в два раза (например, с десяти до двадцати) без увеличения числа посещаемых узлов. При этом, естественно, возникает опасность исключить лучший ход в силу немонотонности оценивающей функции. Например, в шахматах часто возникают позиции, при которых можно отдать ферзя (или другую фигуру), чтобы следующим ходом поставить мат. Отдача ферзя существенно уменьшает значение статической оценивающей функции, так что этот ход, вероятно, окажется худшим по значению оценивающей функции и будет исключен из рассмотрения. Как можно снизить опасность пропуска важной ветви на дереве вариантов?

Для уменьшения риска пропуска хорошего решения могут использоваться процедуры *суживающего n-наилучшего направленного сокращения*, в которых число просматриваемых ветвей *n* в процессе поиска уменьшается по мере продвижения вглубь по дереву вариантов. Закон уменьшения значения *n* с ростом глубины рассматриваемых узлов является эвристическим.

Существует и более изощренный способ уменьшить риск пропуска важных узлов. Можно заметить, что процедура формирования рабочих оценок призвана улучшить стати-

ческую оценивающую функцию за счет перебора. Однако в n-наилучшем направленном сокращении отсечение ветвей ведется на основе значений (неулучшенной) оценивающей функции. Вместо этого можно осуществлять *неглубокий поиск* в целях получения более адекватных оценок для отсечения неперспективных ветвей в более глубоком и узком поиске. Таким образом, сначала осуществляется неглубокий, но широкий поиск для получения предварительных рабочих оценок, уточняющих значения оценивающей функции, затем эти значения используются для осуществления более узкого и более глубокого поиска, который позволяет далее уточнить рабочие оценки для некоторой части отобранных узлов, и т. д. Возможно формирование целой иерархии поисков.

В рассмотренных процедурах направленного сокращения число отсекаемых ветвей, по сути, не зависит от текущей ситуации и уже просмотренной части дерева вариантов. Есть ли здесь резерв для дальнейшего улучшения процедур поиска?

В эвристическом программировании широко используется *процедура альфа-бета-отсечения* (также называемая *методом ветвей и границ*). В этой процедуре промежуточные решения, найденные в процессе перебора, используются для установления порога на значения оценивающей функции, по которому производится отсечение ветвей в еще не рассмотренной части дерева вариантов. Вводится параметр α, который устанавливается в значение качества текущего лучшего найденного решения. Если оптимистическая оценка (т. е. оценка сверху) для некоторого узла меньше значения α, то рассматривать ветвь, выходящую из этого узла, не имеет смысла. Второй параметр, β, используется в игровых задачах для отсечения ветвей, соответствующих ходам противника (т. е. этот параметр характеризует для игрока пессимистическую оценку качества позиции).

В ряде задач можно получить строгие оценки сверху и снизу для решений, достижимых из каждого узла. В этом случае процедура альфа-бета-отсечения позволяет отсекать ветви без опасности пропуска оптимального решения.

Одной из самых известных задач такого рода является *задача коммивояжера*. В этой задаче дается множество городов и указывается стоимость перемещения между каждой парой городов. Требуется найти маршрут, начинающийся и заканчивающийся в одном и том же городе и проходящий через все имеющиеся города по одному разу. Это типичная NP-полная задача. Для нее несложно построить дерево вариантов: каждый узел соответствует последовательности уже посещенных городов, а каждая из выходящих из него ветвей — перемещению в один из новых, еще не посещенных городов. Строгая оптимистическая оценка для некоторого узла складывается из стоимости уже пройденного маршрута и суммы минимальных расстояний до еще непосещенных городов.

С помощью жадного алгоритма можно найти начальное решение: из текущего города мы идем в тот город, до которого стоимость перемещения минимальна среди непосещенных городов, и так поступаем, пока не посетим все города, после чего замыкаем маршрут. Стоимость найденного маршрута может использоваться для отсечения неперспективных ветвей на дереве вариантов: если стоимость начала пути (с учетом минимальной стоимости его завершения) уже больше, чем стоимость текущего найденного лучшего маршрута, то продолжать перебор в этом направлении не имеет смысла. С помощью процедуры альфа-бета-отсечения удается заметно сократить перебор, однако задача все же остается NP-полной.

Для многих задач (особенно игровых) получить точные оценки сверху и снизу для качества некоторой позиции оказывается затруднительно, в связи с этим в процедуре альфа-бета-отсечения могут использоваться и нестрогие (эвристические) оптимистические и пессимистические оценки. В этом случае нет гарантии того, что лучшее решение не будет пропущено, однако процедура альфа-бета-отсечения позволяет выбирать число отсекаемых ветвей более «интеллектуально», чем процедура n-наилучшего направленного сокращения.

На пути дальнейшего усовершенствования процедур формирования рабочих оценок до сих пор сделано немного: ведь эти процедуры лишь перераспределяют число просматриваемых ветвей между узлами. Выбор того, какие именно

ветви просматриваются, зависит от предметно-зависимых эвристик, в частности, от оценивающей функции. При разработке эвристических программ наибольшие усилия уходят именно на разработку частных эвристик, всецело зависящих от конкретной задачи и не переносимых на другие задачи. Никакой теории правильного задания эвристик толком нет. Как результат, в эвристическом программировании наибольший интерес представляют лишь самые общие принципы.

УСПЕХИ И НЕУДАЧИ ЭВРИСТИЧЕСКИХ ПРОГРАММ

Одним из традиционных вызовов области искусственного интеллекта были шахматы. То, что компьютер когда-либо сможет обыграть гроссмейстера, в 1950-е годы верили лишь самые оптимистично настроенные компьютерщики. Даже среди инженеров и ученых, тесно общавшихся с компьютерами, многие считали эту возможность нелепой. Вызывающее название «искусственный интеллект», полученное молодой наукой, порождало заметный скептицизм и откровенную критику, подкрепленную большим обилием псевдонаучных аргументов.

Одним из наиболее известных критиков искусственного интеллекта, резко отрицавших возможность наделения цифровых компьютеров какими бы то ни было человеческими способностями, был философ, профессор Герберт Дрейфус. В своей обширной работе (1965 г.) с весьма показательным названием «Алхимия и искусственный интеллект» (выполненной им в корпорации «RAND») он, в частности, категорично заявлял, что компьютер обладает настолько ограниченными возможностями, что принципиально не сможет играть в шахматы даже на посредственном уровне. Вызов, брошенный Дрейфусом, был принят программистами из лаборатории искусственного интеллекта Массачусетсского технологического института, одним из основателей которой был Марвин Минский.

В 1969 году с Дрейфусом сыграла программа, написанная Ричардом Гринблаттом. Несмотря на уверенную игру профессора и то, что программа исполнялась на компьютере PDP-6 (с быстродействием примерно 0,25 млн операций в

секунду), в результате ожесточенной битвы победа все же осталась за компьютером. Однако этот проигрыш не пошатнул уверенности Дрейфуса в ущербности компьютеров, который и спустя десятилетия продолжал говорить о том, что область ИИ претерпела предвиденную им неудачу. Стоит все же отметить, что критика Дрейфуса содержала не только ряд заблуждений, но и некоторые ценные идеи.

Интересно, что и после этого поражения многие гроссмейстеры говорили о том, что компьютер никогда не сможет играть сильнее любителя. Справедливости ради нужно сказать, что и среди гроссмейстеров были энтузиасты компьютерных шахмат, такие как Михаил Моисеевич Ботвинник, веривший в возможность создания программы, играющей сильнее человека. Все это говорит о том, что критика искусственного интеллекта была не глубже простого оптимизма его сторонников.

Сейчас можно уверенно утверждать, что компьютер способен выиграть в шахматы у чемпиона мира. В отдельных партиях компьютер брал верх уже в 1990-х года, а в 2006 году компьютер выиграл в серии партий у действующего чемпиона мира Владимира Крамника со счетом 2 : 4. Эта участь постигла шашки еще раньше. А в 2008 году на чемпионате по игре в покер (в один и его ограниченных вариантов) компьютер смог победить сильнейших игроков-людей, хотя еще в 2007 году он потерпел поражение. И даже в американской телевикторине «Jeopardy!» в 2011 году компьютер выиграл у сильнейших игроков-людей. И лишь в немногих играх, таких как го, по-прежнему сохраняется безоговорочное лидерство человека. Поклонники все еще не поддавшихся компьютеру игр уже реже говорят о том, что человек навсегда останется в них непобежденным, а чаще просто относят эту страшную дату на многие годы вперед. Стоит, правда, отметить, что прогресс в силе компьютерных игроков для этих игр может быть связан не с вполне классическим эвристическим программированием. К примеру, в го компьютер не так давно поднялся на следующую ступень мастерства благодаря объединению поиска по дереву вариантов с оцениванием по методу Монте-Карло (путем статистических испытаний).

Однако критики искусственного интеллекта не отступают в своих позициях ни на шаг, а просто делают менее конкретные заявления: пусть компьютер выигрывает в шахматы у человека, но он при этом не проявляет творчества, а выполняет лишь то, что заложено в него программистом. Насколько эти «обвинения» справедливы?

Естественно, само умение играть в шахматы или другую игру заложено в компьютер человеком, и за хорошей игрой программы стоит интеллект разработчиков. Однако и человек не достигнет высокого уровня, не переняв огромный мировой опыт игры в шахматы, накопленный на протяжении многих веков. Без объяснений более опытных игроков или хотя бы просто игры с ними любой человек останется на весьма примитивном уровне игры. Хорошо, пусть человек опирается на чужой опыт, но этот опыт все же накоплен людьми. Пусть понемногу, но каждый хороший игрок привносит в эту мировую копилку опыта что-то свое. А компьютер? Он же «механически» выполняет заложенную в него программу, состоящую из строгих операций, которые кто угодно может воспроизвести.

Но ведь и сами шахматы — чисто дедуктивная игра. По теореме о минимаксе оптимальная стратегия игры в шахматы уже содержится в правилах этой игры, поэтому если нам эту стратегию сообщат, мы, согласно классической теории информации, новых сведений не получим. Когда кто-то говорит, что компьютер, выигрывая в шахматы, делает только то, что заложено в него человеком, он, по сути, мыслит в рамках этих классических представлений. Подобные представления берут свое начало в философии нового времени в споре последователей эмпиризма и рационализма о методах познания. При этом индуктивному методу Фрэнсиса Бэкона приписывается возможность получения новых, но обязательно недостоверных знаний, тогда как дедуктивному методу Рене Декарта — достоверных, но принципиально не обладающих новизной знаний. Таким образом, будучи последовательным, следовало бы сказать, что и гроссмейстер своей игрой не показывает ничего, что не было бы заложено в шахматы.

На примере шахмат видно, что этот традиционный взгляд, воплощенный и в современной теории информации, требует

расширения. Еще более ярко это видно на примере великой теоремы Ферма, согласно которой уравнение $a^n + b^n = c^n$ не имеет натуральных (положительных целочисленных) решений при целых значениях $n > 2$. На протяжении нескольких столетий ее пытались доказать многие математики. Порой даже думали, что построение доказательства (или опровержения) этой теоремы — неразрешимая задача («доказательство» этой теоремы среди любителей было настолько же популярным, как и «изобретение» вечного двигателя). Этот случай поучителен и для области искусственного интеллекта: часто говорят о том, что ИИ создать невозможно, поскольку этого не удалось сделать в течение уже более чем полувека, и также проводят аналогии ИИ с вечным двигателем или философским камнем. Однако на доказательство какой-то одной совершенно простой по своей формулировке теоремы Ферма ученые потратили гораздо больше времени, и они бы не достигли успеха, если бы думали о том, как долго ее не удается решить.

Теорема Ферма была истинной независимо от того, что люди об этом не знали. Истинность этой теоремы дедуктивно следует из ее формулировки. Но неужели это должно значить, что доказательство теоремы Ферма, окончательно представленное лишь в 1995 году, не дало нам новой информации о справедливости теоремы? Неужели напряженный труд многих ученых не породил новой информации? А как бы мы отнеслись к этому доказательству, если бы его нашла компьютерная программа (по крайней мере, новые доказательства более простых теорем удавалось получать автоматически)?

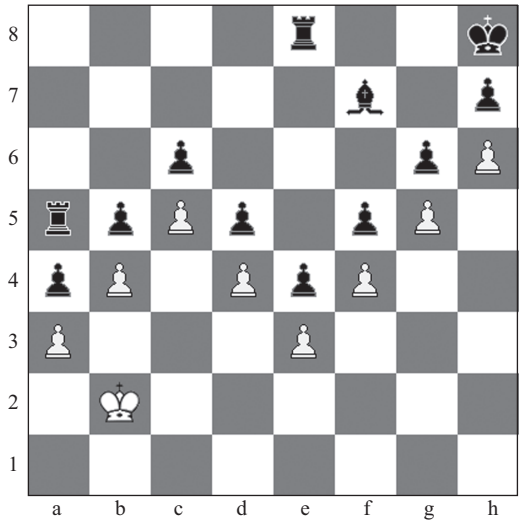
Видимо, необходимо создание теории «дедуктивной информации», в которой количество информации определялось бы через число операций, необходимых для ее получения. Может быть, в рамках такой теории удалось бы определять количество информации, порождаемое некоторой эвристической программой или содержащееся в некоторой эвристике, что дало бы возможность более строго и детально исследовать алгоритмы поиска.

Хотя такой теории нет, все же можно сказать, что компьютер, выигрывающий в шахматы у гроссмейстера, в не-

котором смысле порождает новую информацию. Нельзя уверенно утверждать, что человек способен порождать новую информацию в каком-либо ином смысле, поскольку, возможно, он лишь перерабатывает новую информацию, приходящую извне (и его вклад заключается лишь в переработке, затрате «вычислительных» ресурсов). Вполне возможно, важную роль здесь также играет введение элемента случайности в процедуру поиска. Такой стохастический поиск, находящий каждый раз новые решения, вполне может быть осуществлен и компьютером. О роли случайности мы, однако, поговорим позднее.

Помимо не вполне справедливых упреков в том, что эвристические программы не порождают новой информации, также часто говорят, что у них отсутствует истинное понимание игры. Классическим примером, рассмотренным Джейн Сеймур и Дэвидом Норвудом в статье, опубликованной в 1993 году, служит искусственная игровая ситуация, представленная ниже, на рисунке.

В похожей ситуации за белых было предложено сыграть компьютеру «Deep Thought», который до этого одержал не-



Пример игровой ситуации, в которой компьютерная программа делает ход, приводящий к очевидному проигрышу

сколько побед над гроссмейстерами. Компьютер сделал ход b4–a5, при котором пешка белых съедает ладью черных, приводит к скорому проигрышу белых. Для человека очевидна стратегия белых, которая позволяет вести игру до бесконечности, избегая поражения. При анализе этой ситуации человек, правда, использует понятие связанных областей, которое в шахматах обычно не нужно, т. е. человек и компьютер здесь поставлены не совсем в равные условия.

И все же данный пример показывает определенную ограниченность шахматной программы «Deep Thought». Сильнейшие шахматисты еще могут соперничать с машиной, создавая нестандартные ситуации, смысла которых компьютер не понимает. Для шашек это уже в принципе невозможно, поскольку для этой игры в 2007 году было получено слабое решение — алгоритм, гарантирующий идеальную игру, начиная с начальной, но не с любой произвольной, позиции. Отсюда видно, что программа может идеально играть и без «понимания» игры. Можно предположить, что глубина «понимания» игры связана со способом описания игровой ситуации, который находит отражение в оценивающей функции. Если, к примеру, в оценивающую функцию заложены силы фигур, можно считать, что программа «понимает» значимость той или иной фигуры. То же можно сказать и о позиционном преимуществе, складывающемся из каких-то свойств взаимного расположения фигур. Как уже отмечалось, улучшение оценивающей функции позволяет уменьшать глубину перебора при сохранении уровня игры. Понимание игры у существующих компьютерных программ находится на не слишком высоком уровне, что, однако, компенсируется глубоким перебором за счет существенных вычислительных ресурсов. Такой подход обычно называют методом «грубой силы».

Часто говорят: «Если для решения какой-то проблемы вы используете силу, значит, вы не используете ум». Это высказывание традиционно применялось к военным конфликтам (силовому решению проблемы), которых можно было бы избежать дипломатическим путем. Оно широко применимо и в других областях. Например, хорошо известным примером из химии является сравнение промышленного связывания

атмосферного азота с водородом с использованием установок высокого давления, требующих больших энергетических затрат, и азотфиксацией, выполняемой более «интеллектуальным» образом (без использования грубой силы) многими бактериями в почве.

Видимо, данное высказывание можно распространить не только на любую физическую, но и на вычислительную силу: чем больше вычислительных ресурсов вы используете для решения некоторой проблемы, тем менее интеллектуально вы ее решаете. Отсюда, кстати, видно, что появление квантовых компьютеров или просто сверхмощных компьютеров, способных решать какие-то NP-полные задачи за обозримое время, не преодолеет проблему создания искусственного интеллекта, а лишь завуалирует ее.

Таким образом, критика метода «грубой силы» в определенной степени справедлива. Ведь человек не столько совершает интенсивный перебор, сколько эффективно его избегает, и это позволяет ему неплохо решать очень сложные задачи, тогда как метод «грубой силы» эффективен только для сравнительно простых задач. Однако нельзя утверждать, что ситуации, подобные представленному выше примеру шахматной позиции, доказывают принципиальную неспособность понимания компьютерной программой даже такого формального мира, как мир шахматных партий.

И все же человек при решении задач проявляет ряд принципиальных свойств, которыми обладают и эвристические программы. Это и резкий рост времени нахождения точного решения при увеличении сложности задачи, и поиск приближенного решения для задач большой размерности. Интересно то, что часто отличие человека от компьютера видят в способности ошибаться. Однако очевидно, что при решении NP-полных задач большой размерности компьютеру свойственно ошибаться в не меньшей степени. Конечно, человек часто делает ошибки из-за невнимательности, которые компьютеру как будто не свойственны. Однако это связано с тем, что человек непрерывно решает NP-полную (а может, и неразрешимую) проблему огромной размерности, предоставляемую внешним миром в целом, т. е. живет. Партия в какую-либо игру — лишь маленькая часть этой

проблемы, на которую выделяются ограниченные ресурсы мозга. Если бы компьютерная программа не просто играла в шахматы, но одновременно с этим решала множество иных задач, естественно, в некоторые моменты времени ей приходилось бы отвлекать на них значительные ресурсы, в результате чего глубина ее игры могла резко сокращаться, что выглядело бы, как банальная невнимательность.

Иногда, напротив, люди демонстрируют удивительные способности к решению вычислительно сложных задач. Один из наиболее интересных (в контексте нашего разговора) случаев приведен в книге нейропсихолога Оливера Сакса «Человек, который принял жену за шляпу». Сакс описывает свои наблюдения за двумя близнецами, с детства содержавшимися в лечебных учреждениях с диагнозами «аутизм» и «умственная отсталость». Он обратил внимание на игру, в которую близнецы играли друг с другом. В этой игре они по очереди называли шестизначные числа. Как оказалось, все называемые ими числа были простыми. Сакс, решив проверить, вспоминают ли они известные им числа или каждый раз называют новые, нашел в справочниках девятизначные простые числа и, выписав их, попытался принять участие в игре. Близнецы не только с радостью приняли его в игру, сумев удостовериться в простоте называемых им чисел, но стали постепенно увеличивать порядок чисел вплоть до двенадцатого, для которого простых чисел в справочниках того времени не оказалось.

Сакс указывает на то, что алгоритмы поиска двенадцатизначных простых чисел, такие как решето Эратосфена, требуют огромных объемов памяти, которые в момент написания им книги просто отсутствовали у компьютеров. Кроме того, близнецы весьма плохо владели арифметикой и даже сложение небольших чисел выполняли с ошибками. При этом в процессе решения задач глаза близнецов двигались определенным образом, как будто они рассматривали некую карту или сложный зрительный образ. Как эти факты следовало бы интерпретировать?..

Сакс делает вывод о том, что близнецы пользовались не вычислительными или алгоритмическими, а некими иконическими, или визуальными, приемами для поиска и проверки простых чисел. Действительно, способность близнецов

в уме находить двенадцатизначные простые числа кажется поразительной, особенно с учетом того, что данная проблема является NP-полной. А ведь близнецы определенно не могли в уме проверять делимость некоторого числа на все возможные делители. Подумайте только: перед вами число из двенадцати цифр, и вам нужно проверить его делимость на множество других чисел. Сколько времени у вас это займет даже с использованием калькулятора?! Подобные истории хочется отбросить как миф, нелепицу, которая не может иметь места в действительности. Или следует признать, что в мозгу действуют какие-то хитрые неалгоритмические процессы, которые, если только удастся их задействовать, открывают поистине сказочные возможности? Попробуем все же не отмахиваться от фактов, но, в то же время, и не списывать их на необъясненные способности.

Во-первых, для проверки простоты некоторого числа достаточно проверить делимость на все простые числа от двух до корня из анализируемого числа. Для двенадцатизначного числа потребуется порядка миллиона операций без какой-либо дополнительной памяти. Во-вторых, близнецы называли по одному числу за раз, для чего не нужно применять алгоритмы типа решета Эратосфена, находящие все простые числа вплоть до заданного. Есть приемы, позволяющие получать числа, подозрительные на простоту. Такими числами, к примеру, являются числа вида $2^n - 1$ (например, 7, 31 и др.). Можно придумать много других приемов, например произведения простых чисел, к которым прибавлена единица или суммы простых чисел также часто оказываются простыми ($2 \cdot 3 + 1 = 7$, $2 \cdot 3 \cdot 5 + 1 = 31$, $2 \cdot 11 + 3 + 5 + 7 = 37$ и т. д.). Естественно, близнецы не могли пользоваться на уровне сознания явным поиском делителей некоторого числа или решетом Эратосфена. Но здесь важно то, что для реализации их способностей принципиально хватает ресурсов мозга или компьютеров в рамках алгоритмических процессов. Важно также и то, что время поиска и проверки нового числа у близнецов существенно возросло при незначительном увеличении разрядности чисел. Вопрос состоит в том, какова может быть структура этих процессов поиска. Сакс же противопоставляет структуру процессов и их основу,

т. е. путает уровни представления, причиной чего, видимо, является крайне узкое понимание им алгоритмов как непосредственных вычислений. «Иконические» (в его терминологии) процессы могут быть вполне алгоритмическими.

Можно рискнуть предположить, что близнецы играли в «простые числа» сродни тому, как другие люди играют в шахматы или го. Ведь эти игры тоже представляют собой NP-полные задачи, причем гораздо большей размерности. При игре человек не проверяет на сознательном уровне все возможные ходы и все возможные ответы на них. Игрок постепенно накапливает опыт, учится распознавать типичные позиции, сохраняя в памяти огромные базы игровых ситуаций, подсознательно вырабатывает эвристики, позволяющие отсеивать подавляющее большинство неперспективных ходов. Игровая ситуация по своей структуре, не дающейся непосредственно в восприятии, гораздо сложнее даже двенадцатизначного числа. Глядя на доску, опытный игрок тут же видит пару хороших ходов и возможное развитие ситуации, что является гораздо более удивительным (но и более привычным), чем умение искать простые числа.

Вероятно, как гроссмейстеры достигают высокого уровня игры, посвящая ей свою жизнь, так и мастерство близнецов в игре в «простые числа» — результат длительной практики, в ходе которой их мозгом вырабатывались многочисленные эвристики. Конечно, когда говорится об использовании эвристик человеком, вовсе не имеется в виду, что они в мозгу хранятся в явном виде. Речь лишь идет о тех принципах решения NP-полных проблем, которые могут быть общими для человека и компьютера.

Удивительно не то, что близнецы могли находить простые числа весьма высоких порядков, а то, что человек способен освоить в какой-то степени любую совершенно неожиданную NP-полную задачу, которая в естественных условиях перед ним вряд ли бы возникла. Именно это свойство, в первую очередь, отличает человеческий разум от традиционных эвристических программ, которые сами не могут играть в новую, пусть даже самую простую, игру не зависимо от того, в какое количество игр они уже умеют играть. Вряд ли подобные системы можно назвать действительно интеллектуальными.

Основу поиска хода в программах, играющих в шахматы, обычно составляет какой-либо вариант алгоритма альфа-бета-отсечения со сложной оценивающей функцией. Однако помимо этого в программу закладывается большая база дебютов, а также специальные алгоритмы для эндшпилей. По сути, такая программа почти целиком состоит из предметно-специфичных знаний (относящихся к области шахмат), которые и обеспечивают силу игры (в дополнение к глубокому перебору). Многие «знания» жестко кодируются в теле программы, подобно тому, как нами было сделано в примере с игрой «крестики-нолики» 3×3. И лишь совсем немного кода может быть повторно использовано при создании компьютерной программы для другой игры. Человеческий же интеллект отличается универсальностью: один и тот же человек может играть в шахматы, шашки, писать стихи и даже программировать компьютерных игроков.

На самом деле, сейчас и специалисты в области ИИ не рассматривают последние шахматные программы в качестве систем искусственного интеллекта. Конечно, эти программы используют некоторые технологии ИИ, но они разрабатывались как узкоспециализированные программы, предназначенные не столько для продвижения теории ИИ, сколько для демонстрации возможности победы компьютера в интеллектуальных играх, в чем подавляющее большинство людей исходно сомневалось. В середине же XX века неспособность эвристических программ решать новые задачи взывала к осуществлению попыток создания универсальных средств решения интеллектуальных задач.

ОБЩИЙ РЕШАТЕЛЬ ЗАДАЧ

Чтобы не ограничивать возможности компьютера по решению новых задач, необходимо избежать внесения в программу, выполняющую поиск решения, какой-либо предметной информации. В этом случае, однако, на вход программе оказывается необходимым подавать не только условия частной задачи, но также и сведения о предметной области, представленные в понятной программе форме.

Одна из первых и наиболее известных программ, в которой такой подход был реализован, — это «Общий решатель задач» (General Problem Solver — GPS). Первая версия GPS была разработана Гербертом Саймоном и Аланом Ньюэллом (при участии Кристофера Шоу) в 1957 году (развитие и исследование программы осуществлялось до 1969 года) на основе ранее разработанной ими программы «Логик-теоретик». Кроме того, разработка GPS тесно связана с теми исследованиями Саймона по процессам принятия решений в экономических организациях, за которые он получил Нобелевскую премию по экономике 1978 года. Таким образом, GPS рассматривался не просто как компьютерная программа, а как модель человеческих рассуждений. И по сей день упоминания о GPS можно встретить не только в книгах по искусственному интеллекту, но также и по когнитивной психологии. Кроме того, именно работа над GPS привела к формулировке Ньюэллом и Саймоном уже упоминавшейся гипотезы физической символьной системы о реализуемости мышления на произвольных физических носителях, выполняющих символьные вычисления.

GPS, основанный на принципах эвристического программирования, решает задачи поиска цепочек допустимых действий, приводящих заданную начальную ситуацию к желаемой конечной ситуации (цели). Каждая ситуация описывается как совокупность объектов, находящихся в определенных состояниях. То есть задача для GPS описывается как совокупность имеющихся и желаемых состояний набора объектов.

Описание же проблемной среды, также подаваемое на вход GPS, включает описание допустимых действий (операторов) с указанием того, как меняются состояния объектов при выполнении тех или иных действий. Таким способом могут описываться, например, правила перемещения фигур в шахматах, правила интегрирования и т. д. Поскольку состояния объектов могут иметь некоторую внутреннюю структуру, в GPS проблемная среда также включает описание различий между объектами, по которому GPS имеет возможность определять, чем именно желаемое состояние объектов отличается от их текущего состояния.

Как видно, подход, использованный в GPS, полностью соответствует идее Торндайка о природе мышления животных как решения задач с использованием перебора имеющихся в распоряжении операций. Но мог ли GPS решать задачи, аналогичные тем, что Кёлер предъявлял обезьянам? Задача «обезьяна и банан» была одной из первых, на которой тестировался GPS.

В простейшем варианте этой задачи легко можно выделить несколько объектов: ОБЕЗЬЯНА, БАНАН, ЯЩИК, — каждый из которых может находиться в нескольких состояниях (положениях из множества «место-1», «место-2», «под бананом» для обезьяны и ящика и «на полке», «в руке» для банана). У обезьяны есть несколько операций: ИДТИ, ТОЛКАТЬ ЯЩИК, БРАТЬ БАНАН. Для каждого из действий описывается его эффект в зависимости от положений объектов. Условие задачи задается в виде начальных положений объектов и конечной цели, включающей только положение банана «в руке».

GPS легко находит цепочку операций, позволяющих обезьяне заполучить банан. Кстати, можно ли теперь на основе модели эвристического программирования дать хотя бы некоторые содержательные ответы на вопросы об особенностях решения задач обезьянами, поставленные выше, при обсуждении мышления животных? Понятно, что основные особенности процесса решения задач обезьянами связаны с комбинаторным взрывом и эвристиками, используемыми для его преодоления. При увеличении числа действий в цепочке, позволяющей решить задачу, происходит экспоненциальный рост пространства поиска, поэтому, начиная с некоторого момента, даже небольшое усложнение задачи делает ее непреодолимой для обезьяны. Однако если животное успело познакомиться с некоторыми промежуточными решениями, то решение целой задачи может быть описано более короткой цепочкой действий, и обезьяна с ней справится. При этом используются и общие эвристики, к примеру, в перебор не вовлекаются действия над объектами, находящимися вне поля зрения. Подобные эвристики очень эффективно отсекают неперспективные варианты, но не являются универсальными. Человек, в отличие от животных,

может отказаться от подобных эвристик, хотя по умолчанию их использует. Нередко мы говорим: «не подумал», — когда какой-то из вариантов действий выпадает из нашего поля зрения. Решается ли проблема компромисса между вычислительными затратами и полнотой поиска в GPS?

Для решения задач данная программа использует общий метод, называемый *анализом целей и средств*. Этот метод заключается в направленном построении дерева целей, где подцели выбираются таким образом, чтобы уменьшить различие между имеющейся и желаемой ситуацией. В общем случае с каждой подцелью связывается некоторая величина, характеризующая трудность в ее достижении, которая определяется по трудности устранения соответствующих различий. Подход на основе дерева целей не отличается принципиально от подхода на основе дерева вариантов. Основная разница заключается лишь в том, что служит корнем дерева — начальная позиция или желаемая цель.

Стоит отметить, что описание проблемной среды для GPS включало дополнительную информацию, которая, по сути, задавала предметно-специфичные эвристики поиска. Сюда относится таблица связей, которая содержит информацию о том, какие операторы могут использоваться для уменьшения каких различий, а также информация об упорядочении различий по степени значимости (или трудности их устранения). Хотя GPS и требует задания эвристик, они являются внешними по отношению к программе, в связи с чем применение GPS к новой предметной области не требует его перепрограммирования.

Помимо задачи «обезьяна и банан» GPS был способен решать широкий круг задач. В частности, он успешно справлялся с задачами о Ханойских башнях и семи Кёнигсбергских мостах, а также более серьезными задачами, включающими аналитическое интегрирование, некоторые шахматные и геометрические задачи, задачи планирования и т. д. При этом «обучение» GPS решению задач нового типа можно считать заметно более легким, чем обучение ребенка решению этих же задач.

Вероятно, GPS мог бы быстро справиться и со многими тестами IQ, не требующими знаний о реальном мире (или

при условии, что необходимые знания были бы даны в описании проблемной среды), и показал бы высокий уровень «интеллекта», что, однако, говорит не столько о широких возможностях GPS, сколько об ограниченности тестов IQ. Конечно, GPS решал только корректно поставленные задачи. Интересно, что значительная часть упоминавшейся выше книги Пойа «Математическое открытие» посвящена вопросу формализации задач: ведь школьники начинают свое обучение со словесных задач, в большинстве своем становящихся тривиальными после того, как для них удается найти четкую математическую формулировку.

Относительно алгоритмически неразрешимых проблем мы уже отмечали, что важность правильной постановки задачи сложно переоценить: правильно поставленный вопрос — это уже половина ответа. Этот вывод верен и для разрешимых, но сложных задач. Конечно, нельзя сказать, что правильные постановки могут полностью избавить нас от необходимости рассматривать NP-полные задачи, но отсутствие хорошей формулировки проблемы делает пространство вариантов крайне широким, а поиск в нем — чрезвычайно затруднительным. Это хорошо видно на примере творчества или изобретательства, где «лабиринт вариантов» столь разветвлен, что даже просто выявление всех возможных направлений «движения» невозможно. Здесь эти направления должны не просто перебираться, но генерироваться (и лишь потом отсекаться), за что вполне могут отвечать иные механизмы (и даже возможно, что за механизмы порождения и отсечения вариантов отвечают разные полушария мозга). Что уж говорить о проблеме выбора жизненного пути, где мы зачастую просто не рассматриваем никаких возможностей, выходящих за рамки наших случайных стереотипов? Эти стереотипы — эвристики, позволяющие беспощадно отсекать огромное количество неперспективных вариантов. Без них человеческий мозг потонул бы в море вариантов и не нашел бы никакого приемлемого решения. Но вместе с тем стереотипы приводят к пропуску неожиданных решений, которые могут оказаться гораздо более эффективными. И даже в науке большинство ученых сейчас занимается поиском в актуальных направлениях, определенных в текущей на-

учной парадигме, а не открытием новых областей поиска. Конечно, этот подход является оправданным, но все же уж очень он напоминает перебор методом «грубой силы». Умение управлять эвристиками поиска сейчас является своего рода искусством, которому толком нигде не учат, а проявление этого умения рассматривается как признак одаренности (но именно обычные люди с «дефектом» отсутствия гениальности могут превратить ее из природного дара в доступный для каждого инструмент).

Не удивительно, что задачи, требующие внимательного выбора пространства поиска, рассматриваются в теории решения изобретательских задач. Классической задачей этого типа является задача о построении с помощью шести спичек одновременно четырех треугольников. В большинстве своем люди, не знакомые с этой задачей, включают в пространство поиска плоские фигуры, среди которых не находят ответа. Выход в иное пространство поиска требует заметных усилий. Интересно, что разломать спички (если о запрете на это действие забыли сказать в условии задачи) догадаться гораздо проще, чем построить из них пирамиду. В другой классической задаче такого типа необходимо соединить вершины квадрата ломаной линией, состоящей из минимально возможного числа прямых отрезков. Но как реализовать автоматический выбор компьютером пространства поиска?

Вполне естественно, что вопрос о формализации задач был проигнорирован в начале исследований в области ИИ, поскольку к нему при имеющемся тогда уровне знаний было просто невозможно подступиться. Попробуйте представить, какой должна быть программа, способная решать такие простые задачи: «Один кран наполняет бочку за 10 минут, второй — за 12, третий — за 15. За какое время краны наполнят бочку, работая одновременно?» и «Есть три работника, каждый из которых по отдельности может справиться с некоторой работой за 10, 12 и 15 часов соответственно. Сколько времени им необходимо, чтобы выполнить ее вместе?» Совпадают ли эти задачи по содержанию? Какие упрощения нужно сделать, чтобы суметь их однозначно решить? Как до этих упрощений можно догадаться?

По сути, формализация задачи — это построение «лабиринта», который в эвристических программах считался четко заданным. GPS стал первой программой, в которой этот «лабиринт» задавался извне, а не кодировался в теле программы. GPS оказался заметно слабее, чем специализированные эвристические программы: для решения сложных задач ему не хватало сложных эвристик и оценивающих функций, которые можно было задать в виде программного кода, но которые не могли быть выражены на языке описания предметной среды. А без таких эвристик универсальные процедуры поиска оказывались бессильными перед задачами, в которых размеры деревьев вариантов неизмеримо превосходят возможности плохо направленного поиска. В частности, GPS не мог играть в шахматы, поскольку в них не удавалось разбить конечную цель (выигрыш в партии) на непосредственно достижимые подцели.

Хотя GPS не суждено было превратиться в воплощение искусственного интеллекта, он стал переходным звеном между эпохой эвристического программирования и эпохой представления знаний в ИИ. Несмотря на то, что язык описания проблемной среды в GPS оказался сравнительно бедным, и некоторые задачи не поддавались формализации в его рамках, он стал первым языком представления знаний. При обсуждении шахматных программ, использующих метод грубой силы, было отмечено, что более глубокое понимание игры человеком может быть связано с использованием более сложного способа описания модели игрового мира. Да и насколько универсальным вообще является представление процесса решения задач в виде поиска по дереву вариантов?

ЛАБИРИНТ АЛГОРИТМОВ

Как мы видели, созданию универсальных процедур решения задач препятствует необходимость введения предметнозависимых эвристик и задания самого дерева вариантов. Но для всех ли задач такое представление процесса решения вообще является подходящим? Возьмем в качестве примера задачу поиска оптимального пути, с которой во многом и

связана лабиринтная гипотеза мышления. Рассмотрим эту задачу в следующей постановке.

Пусть дана доска $M \times N$ клеток. Для каждой клетки указана стоимость ее посещения: $c_{i,\,j}$, $i = 1, \ldots, M$, $j = 1, \ldots, N$. Требуется найти путь с минимальной стоимостью из клетки с координатами (1, 1) в клетку с координатами (M, N). Здесь путем считается последовательность клеток, в которой каждая последующая клетка примыкает к предыдущей.

Можно ли подойти к этой задаче с позиций эвристического программирования? Как будет выглядеть дерево вариантов? Корень дерева соответствует начальной клетке, а из каждого узла дерева исходят ветви, соответствующие всем возможным перемещениям из текущей клетки в какую-либо соседнюю с ней клетку. Число листьев на дереве вариантов равно числу всех возможных путей от начальной клетки до конечной. Без дополнительных ограничений число путей бесконечно. Можно ввести эвристику: рассматривать только пути, в которых любая клетка посещается не более одного раза.

Даже с учетом введенного ограничения число возможных путей растет чрезвычайно быстро с ростом размера доски, что делает невозможным их полный перебор уже для сравнительно небольших значений M и N. Можно было бы использовать алгоритм направленного сокращения, который просматривает только некоторые наиболее перспективные направления. Этот алгоритм применим на практике, но получение оптимального решения с его использованием вовсе не гарантировано.

В то же время поиск наилучшего пути вовсе не требует полного перебора. Хорошо известен следующий простой алгоритм:

1) создать массив **D** размера $M \times N$ и установить его элементы $d_{i,\,j}$ в значение бесконечность $(+\infty)$;

2) изменить значение элемента (1, 1) на $c_{1,\,1}$;

3) для каждого элемента (i, j), значение которого было изменено на предыдущем шаге алгоритма, просмотреть его соседей и сравнить значения $d_{i,j} + c_{i\pm1,\,j\pm1}$ с соответствующими текущими значениями $d_{i\pm1,\,j\pm1}$, т. е. определить, будет ли перемещение из текущей клетки в соседнюю обладать

меньшей стоимостью, чем ранее найденный оптимальный маршрут (если таковой уже был). В случае, если $d_{i\pm1,\ j\pm1} > d_{i,\ j} + c_{i\pm1,\ j\pm1}$, заменить значение элемента $(i\pm1,\ j\pm1)$ на $d_{i,j} + c_{i\pm1,\ j\pm1}$;

4) повторять шаг 3 алгоритма до тех пор, пока значения каких-либо элементов массива **D** меняются;

5) массив **D** будет содержать минимальные стоимости путей от точки (1, 1) до каждой из точек доски, откуда будет несложно восстановить и сам маршрут от начальной до конечной точки; при этом время работы алгоритма линейно зависит от площади доски.

Этот алгоритм проиллюстрирован ниже.

За счет чего описанное решение оказывается столь эффективным по сравнению с полным перебором? Данная задача поиска пути перестает быть NP-полной, поскольку при посещении клетки нам не важно, каким путем мы в нее пришли, важна лишь его стоимость. Иными словами, задача разбивается на независимые подзадачи, так как длина последующего пути из некоторой клетки не зависит от уже пройденного пути до нее. Для человека это кажется вполне очевидным предположением, но оно не возникает само по себе в компьютерной программе, а должно быть специально запрограммировано.

Свойство независимости фрагментов решения встречается во многих задачах, однако даже в некоторых разновидностях задачи поиска пути оно может нарушаться. К примеру, при

1	3	2	6
7	6	1	8
4	2	1	5
6	1	8	6
7	2	4	5

1	4	6	12
8	10	7	15*
12	12	8*	+∞
18	13*	+∞	+∞
25*	+∞	+∞	+∞

1	4	6	12
8	10	7	15
12	10*	8	13*
18	13	16*	+∞
25	15*	+∞	+∞

Поиск кратчайшего пути: *а* — массив стоимостей $c_{i,j}$ посещения клеток; *б* — массив D после пяти итераций; *в* — массив D после шести итераций. Отмечены клетки, значения $d_{i,\ j}$ в которых менялись на предыдущей итерации

перемещении сразу нескольких интеллектуальных агентов, которые могут преграждать друг другу путь, стоимость посещения каждой клетки будет меняться со временем, вернее, она будет зависеть от выбора траекторий самими агентами. Такая задача называется задачей *отслеживания пути*, которая оказывается несравненно более сложной и уже не вполне формализуемой.

Если в какой-то задаче поиск фрагментов решения может выполняться независимым образом, то, как правило, существует алгоритм класса P, позволяющий решать эту задачу при произвольных данных. Но даже если независимость фрагментов решения не полная, то их разделение может использоваться в качестве не строгого приема построения алгоритма класса P, а эвристики, позволяющей сократить перебор при поиске приближенного решения. В качестве примера можно представить себе задачу о коммивояжере, который должен посетить все крупные города на нескольких континентах. Естественным было бы искать оптимальный путь по каждому континенту в отдельности, выделив тем самым слабо связанные компоненты задачи. Связь между подзадачами заключается, по крайней мере, в том, что начальный город, в котором начинается движение по новому континенту, должен быть одним из ближайших к конечному городу на предыдущем континенте. Кроме того, нет гарантии, что единичное посещение каждого из континентов даст действительно оптимальное решение.

Таким образом, независимость частей решения некоторой задачи можно считать общей эвристикой, которую можно было бы попробовать заложить в некоторый универсальный алгоритм поиска, однако этого будет недостаточно в других задачах. На примере вопроса о разделении задачи на подзадачи также прослеживается связь между проблемой поиска общих алгоритмов решения задач класса P и поиска эвристик, используемых при решении NP-полных задач. Человек, начав играть в некоторую игру, о существовании явной выигрышной стратегии в которой он не осведомлен, сначала будет выявлять нестрогие эвристики, постепенно повышая эффективность перебора, пока, наконец, эти эвристики не превратятся в точную выигрышную стратегию.

Интересно, что даже крысы в лабиринте не просто выполняют поиск в нем, но осуществляют перебор разных стратегий поиска. В книге К. Э. Фабри «Основы зоопсихологии» приводится пример, когда некоторые крысы сначала сворачивали на всех развилках в одну и ту же сторону. Убедившись, что эта стратегия не работает (при поиске кормушки или выхода), они начинали сворачивать всегда в другую сторону или начинали чередовать повороты направо и налево. Итак, перебор алгоритмов поиска вовсе не редок.

Описанный выше алгоритм поиска пути заметно быстрее полного перебора всех возможных путей, но и он может оказаться неэффективным. Представим, что размеры клеток на доске уменьшаются, а их количество увеличивается, пока не становится чрезвычайно большим. Ситуация становится близкой к реальному миру, в котором из некоторой точки можно попасть в другую точку по произвольной траектории. Осуществлять поиск пути приведенным алгоритмом (не говоря уже об универсальных алгоритмах поиска на дереве вариантов) в таком пространстве становится практически невозможным.

Обратите внимание на то, что в задаче «Обезьяна и банан», предлагавшейся GPS, возможные положения объектов задаются дискретно. Каких бы размеров стало пространство поиска, если бы число возможных положений было сильно увеличено? Какие эвристики потребовалось бы задать, чтобы эта простая задача решалась с помощью поиска по дереву вариантов?

Даже более простая задача — поиск максимума некоторой вещественной функции от нескольких аргументов — вряд ли может быть решена с помощью методов эвристического программирования. Для ее решения оказывается необходимым применять некоторые дополнительные предположения, которые также можно назвать эвристиками. Наиболее часто применяемым предположением здесь является предположение о непрерывности, согласно которому малые изменения значений аргументов функции приводят к малому изменению значений самой функции. Как результат, оказываются применимыми такие методы, как градиентный подъем (или градиентный спуск в случае поиска минимума), в котором

происходит итеративное перемещение из начальной точки в последующие точки небольшими шагами по направлению градиента, указывающего направление наискорейшего локального возрастания функции.

Выше уже упоминалось сходство жадных алгоритмов с подъемом в гору при плохой видимости. По сути, градиентный подъем воплощает такой подъем в гору (иногда он так и называется), т. е. является вариантом жадного алгоритма. Однако при градиентном подъеме не просматриваются все соседние точки для определения оптимального направления, а оценивается значение градиента, и, что еще важнее, перемещение в следующую точку происходит на конечное расстояние, т. е. все промежуточные точки между текущей и следующей точками просто пропускаются (в силу предположения непрерывности полагается, что значения функции в этих точках находятся между значениями в текущей и следующей точках).

Итак, для проблем поиска в непрерывных пространствах в определенной степени подходит терминология эвристического программирования, однако используемые алгоритмы обладают заметными особенностями.

Эвристика непрерывности, как и эвристика независимости, является весьма общей. Можно было бы подумать: давайте под каждый обширный класс задач создавать свои эвристические алгоритмы (в принципе, именно так обычно и поступают), поскольку общих эвристик не так много. Однако при решении новых конкретных задач все равно бы пришлось вводить все больше частные эвристики, так как общих эвристик было бы недостаточно. Особенно это хорошо заметно на задачах класса P.

Рассмотрим, к примеру, следующую игру. Есть кучка из 2012 монеток. Два игрока по очереди берут из нее от одной до шести монеток. Выигрывает тот, кто делает последний ход.

Здесь вполне очевидным образом может быть сформировано дерево игры: из узла, соответствующего текущей ситуации, выходит 6 ветвей, соответствующих взятию от 1 до 6 монет. Размер дерева игры при этом оказывается более чем 6^{235}. Какие эвристики могут быть здесь использованы для сокращения такого чудовищного числа вариантов? В дей-

ствительности, в данной игре имеется явная выигрышная стратегия. Первым ходом игрок должен взять столько монеток, чтобы оставшееся число делилось нацело на 7. После этого, чтобы выиграть, достаточно брать $7 - k$ монеток (где k — число монеток, взятых соперником при последнем ходе). Как видно, для игры в эту игру никакой перебор не нужен. Правда, при известной выигрышной стратегии игра теряет какую бы то ни было интеллектуальность.

Интеллект здесь проявляется не в игре по известной выигрышной стратегии, а в поиске этой стратегии, которая является относительно простым алгоритмом. Эта задача уже будет требовать перебора вариантов в пространстве алгоритмов, число которых растет очень быстро с длиной рассматриваемых алгоритмов, т. е. эта задача является NP-полной (а порой и вообще алгоритмически неразрешимой вследствие проблемы останова). Однако такие задачи встречаются достаточно часто и неплохо решаются человеком. Попробуйте, к примеру, решить следующую задачу, для которой также можно было бы построить дерево вариантов: можно ли замостить доску 8×8 клеток с двумя вырезанными уголками (A1 и H8), т. е. доску из 62 клеток, костяшками домино 1×2?

Для таких задач хорошо видно, что поиск производится не в пространстве состояний игрового мира (в последнем случае не нужно производить перебор возможных вариантов замощения). Вместо этого перебираются разные представления (способы описания) этого мира. В последней задаче достаточно заметить, что на доске черных и белых клеток разное число (30 и 32), в то время как каждая костяшка покрывает по одной белой и черной клетке, поэтому замощение невозможно. Нанесение какой-то раскраски на доску — это введение новых элементов описания, существенно облегчающих процесс решения. Интересно, что задачи, являющиеся NP-полными в рамках одних представлений, оказываются принадлежащими классу P в рамках других представлений (при этом, однако, сам переход между описаниями задачи в рамках разных представлений оказывается NP-полной проблемой).

Процесс построения общего решения задачи класса P также является процессом поиска, однако методы эвристического программирования в организации такого поиска пока помогают мало. Можно рискнуть высказать следующую гипотезу: поиск явного алгоритма решения задач класса P аналогичен построению эвристик в случае NP-полных задач, причем эвристики в общем случае должны представляться в виде алгоритмов.

Видно, что, закладывая в компьютер алгоритм решения некоторой задачи класса P, мы не добавляем ему интеллекта. Наделение компьютера большим числом эвристик для решения какой-то NP-полной задачи также приводит к созданию малоинтеллектуальных программ. Хотя в последнем случае компьютер проявляет некоторые признаки мышления, но это мышление «игрушечное» — на уровне тестов IQ или решения элементарных головоломок, пасьянсов и т. д., служащих «жвачкой для мозга», удовлетворяющих его потребность в мыслительной деятельности, но на «непитательном» материале.

Ранее уже отмечалось, что часто эвристические программы вполне справедливо обвиняют в отсутствии интеллекта в связи с тем, что они не способны решать новые для них задачи. Из этого, однако, делают вывод, что компьютер в принципе не способен делать ничего нового. В действительности, если компьютер будет осуществлять поиск в алгоритмически полном пространстве, то он сможет находить решения совершенно новых для него задач. Возможно, именно эта способность и нужна для сильного ИИ. Поиск алгоритма общего решения некоторой задачи класса P, равно как и поиск новых эвристик для NP-полных задач, — практически неизученные проблемы. Одной из немногих попыток построения программы, которая бы могла самостоятельно выбирать эвристики, была программа «Эвриско» Дугласа Лената, хотя и ее успех был весьма ограниченный. Но пока для нас важна принципиальная осуществимость такого поиска, чтобы раньше времени не впадать в отчаянье от кажущейся ограниченности компьютера.

Итак, поиск, выполняемый некими переборными алгоритмами, является хорошим кандидатом на роль сущности мышления или, по крайней мере, какой-то значимой его компоненты. Однако механизмы подобного перебора до сих

пор остаются загадкой. При попытке ее разрешения выявился целый комплекс проблем, и стало ясно, что поиск — это лишь вершина величественного айсберга мышления.

Путем автоматического поиска в лабиринте алгоритмов не шли из-за его непрактичности: он не позволял сразу разрабатывать системы, пригодные для использования при решении прикладных задач, что стало требоваться в 1970-х годах от области ИИ для подтверждения ее права на существование. В GPS эвристики были представлены не в виде алгоритмов. Напротив, они были вынесены за пределы программы. С одной стороны, это являлось одной из причин ограниченности GPS при решении сложных задач, но, с другой стороны, позволяло ему решать задачи из совершенно разных областей. На первый план выступила проблема представления знаний о предметных областях.

Если вернуться к представлению мышления как «свернутому» действию и поиску в «ментальном пространстве» взамен поиска в физическом пространстве, то можно заметить одну крайне важную упущенную деталь: как представляются и хранятся сведения о внешнем мире, позволяющие формировать пространство состояний и устанавливать допустимые операции, т. е. осуществлять поиск без физического выполнения действий? По сути, это вопрос о том, в каком виде представляется модель мира. При рассмотрении проблемы поиска на примере «игрушечных» миров этот вопрос казался не слишком актуальным до тех пор, пока не стали предприниматься попытки реализовать возможность поиска в любом простом мире из некоторого множества миров либо реализовать этот поиск в сложном мире.

Интересно, что перебор подразумевает необходимость создания и хранения в памяти виртуальной копии мира, что может служить объяснением способности человеческого разума создавать в воображении огромные виртуальные миры (например, при чтении художественных книг), практически полностью отделенные от восприятия реальности, но описываемые в рамках тех же представлений. Но, все же, в каком виде можно представлять знания о предметной области при решении задач компьютером?

Часть вторая

ПРЕДСТАВЛЕНИЕ ЗНАНИЙ

ПРЕДСТАВЛЕНИЕ ЗНАНИЙ КАК ЯЗЫК ОПИСАНИЯ МИРА

Часто мы называем умным того человека, который много знает. Как много маленький ребенок задает вопросов обо всем вокруг, инстинктивно нуждаясь в знаниях! Немного утрируя, можно сказать, что мышление совсем без знаний вряд ли возможно, а при их наличии — и не особо нужно. Зная таблицу умножения, человек легко ответит на вопрос, сколько будет шестью девять, тогда как выполнить в уме элементарный алгоритм по перемножению двух трехзначных чисел сможет не каждый. Также и при игре в шахматы гроссмейстер, не тратя времени на раздумья во время сеанса одновременной игры, выиграет за счет своих знаний у десятка новичков, сколько бы они ни перебирали варианты. Выходит, главное для интеллекта — знания? Ведь мы очень часто человеку, который нас поражает своими широкими знаниями, говорим: «Какой ты умный!» С другой стороны, мы уже отмечали, что в тестах на уровень интеллекта вопросы на знания нам кажутся не вполне уместными. Так как же интеллект и знания соотносятся между собой?

Выше уже отмечалось, что для перебора цепочек действий в процессе мышления необходимо обладать моделью мира. В традиционном понимании *модель* объекта — это другой объект (материальный или, что для нас интереснее, информационный), такой, что при независимом применении одних и тех же воздействий на оригинальный объект и модель между ними сохраняется соответствие по некоторым важным

Часть третья

МАШИННОЕ ОБУЧЕНИЕ

ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

ЧТО ЗНАЧИТ УЧИТЬСЯ?

Нередко говорят, что компьютер умеет лишь то, что заложил в него человек, относя это на счет и программ, решающих какие-то задачи лучше человека (например, играющих в шахматы). Отсюда делают вывод о невозможности создания сильного ИИ. Мы уже обсуждали спорность этого тезиса даже относительно программ, от начала и до конца разработанных человеком. А теперь представим себе программу, которая исходно играет в шахматы плохо, но качество ее игры улучшается по мере приобретения опыта. Разве можно сказать, что эта программа умеет делать лишь заложенное разработчиком? Конечно, способность к обучению должна быть «врожденной» (каковой она является и у человека), но умение играть будет приобретено уже самостоятельно. Способность компьютеров к обучению развеяла бы многие мифы об их ограниченных возможностях.

Что же такое обучение, и отчего возникает в нем потребность для интеллекта? Вполне очевидно, что процесс обучения связан с получением новой информации. Это видно не только на примере экспертных систем, но также и из повседневной жизни. Конечно, сложно, не прибегая к мистике, представить себе разум, который бы заранее обладал полной информацией обо всем мире — он должен был бы содержать в себе всю Вселенную или даже превосходить ее. Хотя такая

ситуация не редкость в искусственных замкнутых мирах типа шахмат, но и там возникает потребность в обучении решению задач — поиску таких цепочек собственных действий, которые могут привести мир из текущего в желаемое состояние. Подобное обучение можно трактовать как накопление «дедуктивной» информации.

Однако просто накопление информации еще не означает обучения. Представим себе человека, обучающегося игре в шахматы, который смотрит на игры гроссмейстеров и запоминает ходы, сделанные в той или иной ситуации. Свои знания он сможет применить лишь в абсолютно таких же ситуациях, если они вдруг возникнут (на что, однако, надеяться вряд ли стоит). Или можно представить себе ученика, который запоминает решение каждого конкретного уравнения вместо того, чтобы понять общий принцип нахождения ответов. Подобную «зубрежку» вряд ли можно назвать обучением, разве что самым примитивным, несмотря на то, что при этом происходит запоминание максимального объема информации. Здесь уместно еще раз вспомнить пример из книги Феймана, приведенный выше. Очевидно, самим обучающимся интеллектом должна выполняться дополнительная обработка информации с формированием каких-то общих правил на основе частных примеров.

Но почему тогда просто не запоминать общие правила, уже выведенные кем-то ранее? Разработчики экспертных систем шли именно этим путем и натолкнулись на существенные трудности. Да и при обучении людей запоминание общих правил недостаточно, ведь не зря в образовании имеет столь большое значение дидактика как наука, исследующая закономерности усвоения знаний (результаты дидактики еще, правда, не востребованы при обучении систем, основанных на знаниях, в области искусственного интеллекта). Многие знания экспертов оказываются невербализованными — не выраженными в словесной форме. Эти знания сформированы на подсознательном уровне в результате обобщения большого личного опыта. Об этом свидетельствует большая трудоемкость работы инженеров по знаниям при разработке ЭС: те знания, которые эксперты могут ясно сформулировать, оказываются недостаточными для создания ЭС, и приходится

производить сложную процедуру извлечения знаний — выявления и формализации результатов обобщения индивидуального опыта. Последнее означает, что осознаваемые общие правила — лишь малая часть информации, используемой интеллектом. Особенно ярко это проявляется в «низкоуровневых» задачах, таких как управление движением или распознавание образов. Объяснение мастера единоборств не позволит новичку (или тем более роботу) тут же овладеть каким-то приемом. Также и профессиональный дешифровщик аэрокосмических снимков или умелый игрок на бирже не сможет сформулировать правил, по которым другой человек или компьютер смог бы столь же эффективно решать соответствующую задачу.

Переход от сенсорно регистрируемой ситуации, представленной, например, в форме яркостей пикселей изображения, к ее словесному описанию оказывается настолько сложным, будто эти два уровня описаний разделены так называемой *семантической пропастью*. Если нет перехода от регистрируемой сенсорами ситуации к общим правилам, то эти правила остаются бесполезной абстракцией. Даже в случае с обучением решению уравнений, когда «сенсорные данные» (уравнения) максимально близки к терминам, в которых формулируется сама задача, простое заучивание алгоритма решения даст лишь способность решать конкретный тип уравнений. Например, человек может заучить формулу для корней квадратного уравнения, приобретя от этого ровно столько же, сколько приобретает компьютер, когда в него закладывается программа решения этих уравнений. Видно, что если образование будет сводиться к заучиванию частных фактов и алгоритмов, то не компьютер по своим возможностям будет догонять человека, а человек будет спускаться до текущего уровня компьютеров. Обучение же должно давать нечто гораздо большее.

Ясно, что обучение как-то связано с получением и обобщением информации, но универсальное определение обучению дать крайне затруднительно — слишком уж оно многогранно. В искусственном интеллекте чаще всего используется частная, хотя и достаточно широкая формулировка *обучения по примерам*: дана *обучающая выборка* (набор примеров задач некоторого типа), и требуется научиться решать любые задачи этого типа.

Интересно, что если человеку сообщается готовый алгоритм решения задачи конкретного типа, то это тоже называется обучением, тогда как в случае компьютеров это называется всего лишь программированием. Конечно, разница в том, что человек как-то пытается (правда, не всегда) вписать этот алгоритм в свою систему знаний. Для компьютера же алгоритм превращается в программу, мало связанную с прочими хранящимися в нем программами. И все же программирование можно назвать самым примитивным «обучением» компьютеров. Другие способы обучения, связанные с автоматическим поиском общего решения по частным примерам, можно тоже разделить по полноте информации, представленной в обучающей выборке. Традиционно выделяют *обучение с учителем*, *обучение с подкреплением* и *обучение без учителя*. При обучении с учителем для каждой задачи из обучающей выборки сообщается правильный ответ, а возможно, и путь решения. При обучении с подкреплением «ученику» лишь сообщается, правильно или нет он решил ту или иную задачу. При обучении без учителя даются лишь формулировки задач без ответов. Конечно, можно представить себе какие-то промежуточные или совсем иные случаи задания обучающей информации, однако эти три вида обучения являются наиболее типичными.

Обучение с учителем чаще используется в прикладных задачах, когда от компьютера требуется вполне определенное функционирование. При этом обучение играет роль своего рода настройки программы. Компьютер пытается воспроизвести именно те решения, которые от него хочет человек, даже если они не оптимальны.

При обучении с подкреплением в роли учителя выступает среда. Она не сообщает детального ответа, а лишь указывает на правильность или ошибочность выбранного решения. Например, когда организм ищет пищу, ему сообщается не правильный маршрут (как в обучении с учителем), а лишь то, нашел ли он в итоге пищу или нет, что сопровождается поощрением (удовольствием от еды) или наказанием (голодом).

Обучение без учителя реже используется в прикладных задачах, поскольку поведение компьютера после такого обучения плохо контролируемо. Найденные решения могут выражаться совсем не в тех терминах, которые бы хотелось человеку. К примеру, если компьютеру дать лишь результаты анализов различных больных, возможно, компьютер сможет (используя надлежащий алгоритм) разделить их на группы, в каждой из которых пациенты будут больны одной и той же болезнью, т. е. по анализам сможет выполнять диагностику. Но никакой связи с названиями болезней не будет. Такое обучение больше характерно для автономных систем и часто называется *самообучением*. Обучение без учителя является наиболее сложным, но обладает и наиболее широкими возможностями, ведь именно этот тип обучения позволяет получить что-то действительно новое. По сути, значительную часть научных исследований можно трактовать как обучение без учителя (или, в крайнем случае, как обучение с подкреплением). Откуда возьмется информированный учитель, когда решения не известны никому из людей и лишь природе можно задать вопрос о правильности решения, и то только тогда, когда есть возможность выйти за рамки чистых наблюдений и поставить правильный эксперимент? В этом смысле обычное самообучение людей по книгам часто является не совсем обучением без учителя, поскольку в нем нет лишь человека-учителя, и поиск обучающей информации приходится выполнять самостоятельно (но эта информация может содержать не только примеры задач, но и подробное описание их решения).

Машинное обучение тоже может быть *пассивным*, когда компьютер не выбирает обучающую информацию, и *активным*, когда компьютер сам отбирает (а то и ищет) наиболее интересную для него обучающую информацию. Чаще всего реализуется простейший вариант активного обучения, при котором компьютер лишь указывает, для каких примеров обучающей выборки ему бы хотелось знать правильный ответ. Этот вариант может быть полезен, когда задание правильных ответов требует от разработчиков больших трудозатрат, в связи с чем для большинства примеров ответы исходно могут отсутствовать. Так, робот, который учится

распознавать изображения, не может просить человека подробно описать ему каждый видеокадр и должен выбирать, информация о каких объектах представляет для него больший интерес.

Кроме того, часто выделяют *инкрементное* и *неинкрементное* обучение. В неинкрементном обучении вся обучающая информация предоставляется компьютеру одновременно. В инкрементном же обучении обучающие примеры появляются последовательно, по ходу его работы, и компьютер должен постоянно корректировать результаты обучения, дообучаться.

Неинкрементное обучение, как и обучение с учителем, чаще используется в прикладных задачах, поскольку соответствует предварительной настройке программы. В случае инкрементного обучения поведение компьютера меняется в процессе работы, что уменьшает его предсказуемость и может осложнить работу человека с ним как с инструментом. В то же время инкрементное обучение может сделать поведение компьютера гораздо более гибким, адаптируемым к изменяющимся условиям. Конечно, такой тип обучения (вместе с активным самообучением) необходим для сильного ИИ.

Наиболее сложный и наименее изученный тип обучения — это *метаобучение*, или обучение способности к обучению (именно развитие этой способности раньше считалось основной целью образования). Сложность метаобучения заключается в том, что оно сводится к построению алгоритмов обучения в конкретных областях, для чего необходимо производить поиск в алгоритмически полном пространстве (без чего, видимо, сильный ИИ недостижим).

Однако, к сожалению, даже для наиболее простого неинкрементного пассивного обучения с учителем трудно в общем виде строго поставить саму задачу обучения, не говоря уже о том, чтобы предложить какое-то ее единое решение. Как подойти к проблеме создания обучающихся компьютеров? И может ли вообще машина учиться? Несмотря на отсутствие четкой постановки задачи обучения, существуют подходы, которые многими считаются универсальными способами обучения (а то и самообучения) компьютера. В первую очередь,

речь идет об искусственных нейронных сетях (ИНС), которые широко известны даже среди неспециалистов в области ИИ в качестве чуть ли не единственного способа создания обучаемых компьютеров. Да и среди специалистов такое мнение об ИНС — не редкость. Посмотрим, насколько нейросетевой подход решает проблему машинного обучения.

ЗАГАДКА НЕРВНЫХ КЛЕТОК

Способность человека научиться почти чему угодно еще более поразительна, чем способность компьютера выполнить любой алгоритм. В конце концов, компьютеры преднамеренно создавались универсальными, а вот способность человека научиться играть в шахматы, раскладывать числа на простые сомножители или сочинять музыку может показаться в высшей степени неестественной. И вправду, для выживания в природной среде все эти навыки вряд ли могут пригодиться. Зачем же уметь им учиться? Конечно, у человека нет специфического умения учиться, скажем, играть в шахматы (хотя, возможно, есть такие генетически заложенные умения, как обучение языку). Просто мозг, как и универсальная машина Тьюринга, может выполнять любые алгоритмы (с ограничением, правда, на вычислительные ресурсы). Такое решение для природы также оказалось проще, чем путь постоянного усложнения мозга с добавлением в него все новых и новых частных алгоритмов. Конечно, в случае с человеком речь идет о способности не только выполнить любой алгоритм, но и самостоятельно его сформировать. Видимо, и здесь природе удалось найти какое-то сравнительно универсальное решение, которое пока не известно человеку (хотя оно и работает в его мозгу). Удивительные способности мозга все время обращают на себя внимание специалистов по ИИ. Такое внимание объяснимо, поскольку эти способности никак не удается в полной мере воспроизвести на компьютере.

Обучаемость человека связывают с *пластичностью* мозга, наиболее явно проявляющуюся (если не считать пластичности на макроуровне — регенерации мозга после повреждений) в изменении связей между нейронами при обучении. Конечно, нейроны у многих животных ничем принципиально не отличаются от нейронов мозга человека. Основное отличие состоит в структуре самого мозга, эволюционное развитие которой (при сравнительно малой изменчивости нейронов) и привело к возникновению человеческого интеллекта. Также и совершенствование интеллектуальных компьютерных систем происходит за счет развития их структуры, алгоритмов, в основе которых лежит небольшой набор базовых операций. Однако не только человек, но и животные обладают способностью к обучению, превосходящей даже современные компьютеры с новейшими программами. Может, нейроны обладают какими-то свойствами, делающими их гораздо более пригодными для задач обучения? Посмотрим, что же собой представляют биологические нейроны.

По своей внутренней структуре нейроны не сильно отличаются от прочих клеток организма. Они содержат ядро, в котором присутствуют хромосомы со стандартным для всего организма набором генов (правда, многие гены «работают» только в мозгу), развитый цитоскелет, а также прочие органеллы (например, митохондрии). Само тело нейрона шарообразной формы, но имеет множество отростков, соединяющихся с отростками других нейронов, что и составляет основное внешнее отличие нервных клеток от прочих клеток. Один из этих отростков называется *аксоном*. Он передает выходные сигналы от своего нейрона другим нейронам. Прочие отростки — *дендриты*. Они обычно соединяются с аксонами других нейронов и получают от них входные сигналы для своего нейрона. Схема связей нейрона показана на рисунке.

Дендриты относительно короткие — менее миллиметра, в то время как аксоны могут быть весьма длинными — более метра, что по меркам клеточного мира очень много. Правда, часто их длина тоже составляет доли миллиметра. Дендриты обычно бывают весьма разветвленными, и один дендрит может получать сигналы от многих аксонов. Сами аксоны ветвятся ближе к концу и могут соединяться не только с дендритами, но и с телами других нейронов. Весь нейрон — его

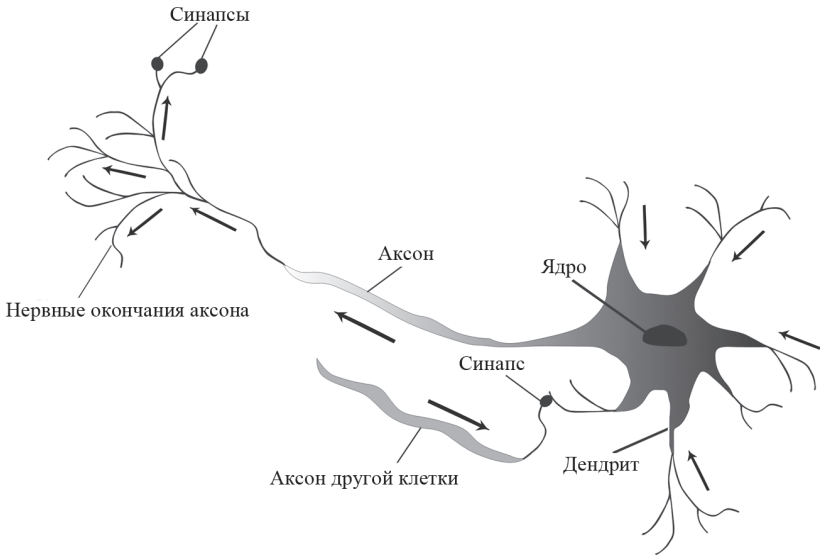


Схема связей нейрона

тело, аксон и дендриты — окутан единой клеточной мембраной, которая и выполняет функции взаимодействия с другими клетками. В местах соединения нейронов их мембраны не соприкасаются. Соединение и взаимодействие клеток осуществляются через *синапс* — особое пространство между двумя мембранами. При этом клетка, от которой по аксону идет сигнал к синапсу, называется *пресинаптической*, а клетка, которая принимает этот сигнал (синапс примыкает к ее дендриту или телу), — *постсинаптической*.

Число клеток в мозгу велико — порядка 10^{12}, — но не больше, чем во многих других внутренних органах, в которых разум почему-то не зародился. Кроме того, нейронов в мозгу лишь около 10^{11}, т. е. всего сто миллиардов, причем в коре головного мозга порядка 10^{10} нейронов. Кроме нейронов в мозге присутствуют *нейроглиальные* (или просто *глиальные*) клетки, которых на порядок больше, чем нейронов (т. е. 10^{12}). Глиа заполняет пространство между нейронами, в связи с чем говорят, что она выполняет опорную функцию: даже само название глии связано со словом «клей» (англ. слово «glue» происходит от греч. «glia»). Кроме того, глиа через межклеточную жидкость осуществляет питание нейронов (так называемая трофическая функция).

Поскольку число клеток в мозге не больше, чем в других органах, сложность мозга лучше видна по числу связей в нем. Ведь каждый нейрон связан в среднем более чем с тысячей других нейронов, что дает 10^{14}–10^{15} связей в мозгу. Насколько это много? Способов реализации этого числа связей очень много: даже каждый отдельный нейрон может установить связи с соседями примерно 10^{300} способами. Число возможных связей между нейронами мозга (как и число комбинаций основ ДНК) часто сравнивают с числом атомов во Вселенной, так же, как и уже упоминавшееся число возможных партий в шахматы. Еще раз необходимо отметить некорректность этого сравнения. Применительно к нейронным сетям это вводит в крайнее заблуждение, будто мозг сложнее Вселенной, что, естественно, совершенно неверно. Ведь и оперативная память компьютера (не говоря уже о прочих накопителях) лишь в 4 гигабайта может находиться свыше чем в $10^{1\,000\,000}$ числе разных состояний — невообразимо большое число, на многие порядки превосходящее физический объем Вселенной, даже выраженный в кубических нанометрах. В действительности, емкость мозга, подсчитанная лишь на основе числа связей, не принципиально превосходит емкость памяти компьютеров (и по разным оценкам составляет сотни или тысячи терабайт). Само по себе, однако, число ячеек памяти не добавляет интеллектуальности компьютеру, как и число клеток — некоторому органу.

Число связей в мозгу (и соответствующее быстродействие) в сравнении с современными компьютерами велико, но обозримо. Принципиальное значение имеет не столько число связей, сколько их структура. Также и более мощный компьютер без программного обеспечения будет полностью бесполезен по сравнению с менее мощным компьютером, но с нужным программным обеспечением. В связи с этим пластичность связей между нейронами (определяющая самостоятельное совершенствование «программного обеспечения» мозга) приобретает особый интерес, и можно ожидать, что где-то в них кроется ответ на загадку обучения и мышления. Тогда нужно обратиться к исследованию не конкретной структуры мозга, а функционирования отдельных нейронов.

Но как работает нейрон? Достаточно давно было замечено, что нейрон может находиться в одном из двух состояний — покоя и возбуждения. В состоянии возбуждения нейрон начинает посылать по аксону нервные импульсы, создающие определенную разность электрических потенциалов. Интересно, что «сила» импульсов не зависит от степени возбуждения клетки, но клетка начинает посылать импульсы только после достижения определенного уровня возбуждения. А уровень возбуждения клетки зависит от уровня возбуждения других клеток, от которых ей поступают импульсы на вход. Немаловажно и то, что есть *тормозные синапсы*: сигнал от пресинаптической клетки, проходящий через такой синапс, не увеличивает уровень возбуждения постсинаптической клетки, а уменьшает его.

Первая известная математическая модель нейрона была предложена МакКаллоком и Питтсом в статье 1943 года (в русском переводе статья появилась в 1956 году под заголовком «Логическое исчисление идей, относящихся к нервной деятельности»). Авторы полагали, что нейрон действует по принципу «все-или-ничего»: он активируется в том случае, если его одновременно возбуждают другие нейроны в количестве, большем заданного порога (взаимное расположение синапсов при этом роли не играет). Если сигнал приходит через хотя бы один тормозной синапс, то нейрон не активируется. Возбужденный нейрон начинает сам передавать сигналы по своему аксону на связанные с ним нейроны. Авторы не рассматривали такие временные характеристики работы нейронов, как, например, скорость распространения нервных импульсов (которая в действительности может варьироваться в диапазоне от 0,1 до 10 м в секунду). Вместо этого предполагалось, что нейроны работают синхронно, по тактам, и информация от каждого возбужденного нейрона передается на следующем такте.

Крайне важно понять, что подобные модельные нейроны способны делать. Как показали авторы, с помощью сети из таких нейронов можно реализовать любую формулу логики высказываний. Кроме того, как чуть позже показал Клини, любой конечный автомат (машину Тьюринга частного типа — без неограниченной памяти) можно воплотить в форме такой

сети, но не более того. С одной стороны, этот результат показывает, что на нейронных структурах могут выполняться алгоритмы частного типа. Значит, связь машины Тьюринга и мышления проявляется не только на уровне доказательства математических теорем, но и на уровне нейронов. Важность этого факта трудно переоценить: ведь он дает ключ, без которого интерпретация нейрофизиологических данных крайне затруднительна. Попробуйте отбросить понятия информации и алгоритмов (которые в середине XX века в их математической формулировке только входили в научный обиход, причем их использование в нейрофизиологии и психологии нередко подвергалось жесткой критике) и представить, как без них будет выглядеть анализ структуры нервной ткани! Между физическим или химическим описанием работы нейрона и его ролью в реализации высших психических функций лежит пропасть, преодоление которой без помощи понятия алгоритма было бы вряд ли возможно. Но, с другой стороны, возникает вопрос, не упущено ли в данной модели что-то важное: ведь нельзя же так просто признать мозг всего лишь большим конечным автоматом?

Вскоре стало ясно, что возбуждающие синапсы должны иметь различную «проводимость», т. е. входы от разных нейронов должны суммироваться с разными весами. Соответствующие принципы открыли Алан Ходжкин и Эндрю Хаксли в 1952 году (за что им совместно с Джоном Эклсом в 1963 году была присуждена Нобелевская премия). Сейчас хорошо известно, что как снаружи, так и внутри нервной клетки, присутствует большое число солей, растворенных в воде и находящихся в диссоциированном состоянии, т. е. в виде ионов, среди которых основными являются ионы хлора, натрия, калия и кальция. Клеточная мембрана нейрона содержит миллионы (задумайтесь над этим числом с учетом общего количества клеток мозга!) «микропор», *клапанов* и *насосов*.

Все поры — это белковые трубки, пронизывающие мембрану (состоящую из жиров), через которые могут поступать и выделяться разные вещества. Это целые механизмы, которые могут либо просто открываться или закрываться, либо прокачивать через себя вещество, затрачивая на это

энергию. Их работа не случайна, а управляется разными факторами. Например, закрытие некоторых пор может осуществляться при возникновении определенной разности потенциалов снаружи и внутри клетки или при поступлении некоторых веществ. При этом поры зачастую пропускают только определенные ионы. К примеру, есть калиевые насосы, натриевые клапаны и т. д.

Обычно насос не просто закачивает внутрь клетки ионы одного вида, но «обменивает» их на ионы другого вида. В частности, на мембране все время работают насосы, которые обменивают ионы калия (попадающие внутрь нейрона) на ионы натрия (выводящиеся наружу). Одновременно с этим открыты калиевые клапаны: через них ионы калия выходят обратно из клетки, стремясь уравновесить концентрацию калия снаружи и внутри. Но как только ионы начинают выходить через клапаны, возникает разность потенциалов: положительных ионов снаружи становится больше, и они препятствуют выходу новых положительно заряженных ионов. Наступает состояние равновесия, в котором снаружи клетки положительно заряженных ионов несколько больше, т. е. имеется разность потенциалов, составляющая примерно 0,07 вольта (говорят, что мембрана *поляризована*). Именно так обстоит дело, когда нейрон находится в состоянии покоя.

Если вдруг в каком-то месте мембраны откроются натриевые клапаны, ионы натрия, выкачанные насосами наружу, устремятся внутрь клетки, тем самым *деполяризуя* мембрану (выравнивая заряды снаружи и внутри клетки). Интересно, что достаточно сильная деполяризация мембраны приводит к дополнительному открытию натриевых клапанов, что приводит к дальнейшему усилению деполяризации мембраны. После открытия всех натриевых клапанов потенциал становится *реверсивным* (заряд внутри оказывается больше, чем снаружи, а разность потенциалов составляет примерно 0,04 вольта). Таким образом, если начальная деполяризация является достаточно сильной, процесс оказывается самоусиливающимся. Он начинает распространяться как волна от области начальной деполяризации (находящейся обычно на теле нейрона) по всей его мембране — вдаль от тела нейрона по аксону. Интересно, что натриевые клапаны быстро

закрываются после своего открытия и некоторое время не способны открываться вновь. В результате в этом месте деполяризация мембраны восстанавливается (этому способствует еще и дополнительное открытие калиевых клапанов), а «волна» реверсивного потенциала может распространяться только в том направлении, где натриевые клапаны еще не открывались. Позади же волны остаются области с только что закрывшимися клапанами. Так формируется нервный импульс, представляющий собой перемещение области деполяризации с разностью потенциалов, не зависящей от величины начальной деполяриации мембраны на теле клетки.

Когда импульс доходит до синапсов, из окончаний аксона в синаптическую щель высвобождаются химические вещества, которые в результате диффузии перемещаются к мембране (дендриту или телу) постсинаптической клетки. Выделение этих веществ, называемых *нейромедиаторами* (существуют десятки разных нейромедиаторов), возникает как реакция на деполяризацию соответствующего окончания. В постсинаптической мембране есть специализированные белковые рецепторы, которые реагируют на соответствующие нейромедиаторы, приводя к открытию тех или иных клапанов. Например, в зависимости от того, будут ли при этом открываться калиевые или натриевые клапаны, соответствующий синапс будет возбуждающим или тормозящим. Кроме типа открываемых пор может меняться и их количество, т. е. синапсы действительно могут обладать разной «проводимостью», которая может меняться.

Открытие небольшого числа клапанов обычно не является достаточным для самоусиливающейся деполяриации мембраны, и возбуждение постсинаптического нейрона остается на подпороговом уровне (а открывшиеся каналы закрываются спустя несколько миллисекунд). Однако, если поступают сигналы сразу через несколько возбуждающих синапсов, открывается большее число клапанов и мембрана может деполяризоваться в достаточной степени, чтобы начались генерироваться нервные импульсы. Иными словами у нейрона есть некоторый «порог срабатывания».

Несмотря на сложность механизмов распространения нервных импульсов, все выглядит так, будто нейрон дей-

ствительно просто суммирует с разными весами входящие сигналы и сам начинает посылать импульсы, если входной уровень сигнала больше некоторого порога. Даже в конце 1980-х годов Нобелевский лауреат Дэвид Хьюбел в своей книге «Глаз. Мозг. Зрение» писал о нейронах, как клетках, суммирующих или интегрирующих поступающую к ним от других нейронов информацию и передающих ее дальше по принципу «все или ничего».

Следует лишь отметить, что выход нейрона нельзя считать бинарным, поскольку возбужденный нейрон генерирует целую серию импульсов. И если «сила» импульсов не зависит от начального уровня возбуждения нейрона, то их частота тем больше, чем сильнее возбуждение. А чем больше частота выходных импульсов, тем больше их суммарное воздействие на постсинаптические нейроны. Правда, нейроны не могут посылать импульсы чаще, чем 1000 раз в секунду, так как в силу закрытия натриевых клапанов примерно на 1 мс наступает так называемый период *рефрактерности*. Заметим, что если число импульсов, проходящих через связи между нейронами мозга, считать за число выполняемых им операций, то производительность мозга составит сотни петафлопсов и будет сопоставимой с производительностью современных суперкомпьютеров (хоть мозг, потребляющий порядка 30 Вт, оказывается заметно более «высокотехнологичным»). При этом большинство ученых полагает несущественным ни распределение импульсов по времени (кроме их числа в среднем), ни распределение их в пространстве, задаваемое взаимным расположением синапсов. В связи с этим учесть частоту импульсов в модели нейрона достаточно просто — нужно лишь заменить *активационную функцию*, которая переводит суммарный сигнал на входе в сигнал на выходе. Вместо ступенчатой функции, равной нулю до некоторого порога и единице — после, активационную функцию можно сделать и более гладкой, но также ограниченной сверху максимальной частотой импульсации.

Так что же, нейрон — это «пороговый сумматор», несмотря на все свое сложное устройство? Конечно, если углубляться в детали работы нейрона, то возникнет впечатление, что математическая модель нейрона к реальности отношения

не имеет. Но в моделях всегда выделяется лишь самое значимое. В конце концов, многие процессы в других клетках тела также сложны, но с мышлением не связаны, поэтому интересны модели нейрона, в которых описываются не все особенности его устройства и работы, а лишь те, которыми он отличается от прочих клеток. Поверим пока, что формальный нейрон воспроизводит, если уж не «внутреннюю жизнь» реального нейрона, то хотя бы его функцию по обработке информации.

Теперь у нас есть искусственные (формальные) нейроны. Осталось взять их побольше, соединить в сеть, которая начнет самообучаться, в результате чего зародится компьютерный разум. Несмотря на наивную абсурдность такой точки зрения, она достаточно распространена, причем не только в фантастике. А действительно, что полезного можно сделать с помощью искусственных нейронов?..

ПЕРЦЕПТРОН

Модель МакКаллока—Питтса стала первым важным шагом на пути развития теории искусственных нейронных сетей. Но, конечно, она не могла не содержать недостатков. Какой же недостаток в первую очередь должен бросаться в глаза? Какое самое важное свойство нейронов отсутствовало в этой модели? Несложно понять, что это свойство — обучаемость. Сейчас надежно установлено, что долговременная память и процессы обучения связаны, в первую очередь, с синаптической пластичностью — изменением связей между нейронами. Однако связи между нейронами у МакКаллока и Питтса считались фиксированными. Как же должны обучаться нейронные сети? По каким принципам у них должны меняться веса связей? И что в итоге обучения должно получаться?

Первую идею о том, как могли бы обучаться нейроны, высказал Дональд Хебб в книге «Организация поведения: нейропсихологическая теория» (1949 г.). В общем виде *правило Хебба* можно сформулировать так: если два нейрона

активируются одновременно, связь между ними должна усиливаться. Это правило интересно тем, что несложно представить себе биологические механизмы его реализации. Каковы цели такого «обучения»? Речь здесь, вероятно, идет не об обучении сети решать какую-то задачу. Можно лишь надеяться, что такая настройка связей приведет к некоторой самоорганизации.

В 1950-х годах Фрэнк Розенблатт предложил обучаемую искусственную нейронную сеть специального вида, названную им *перцептроном* (от perception — восприятие). Эта сеть предлагалась как модель зрительного восприятия. Она должна была учиться распознавать разные объекты по их изображениям. Таким образом, у сети появлялась вполне конкретная задача и, кроме того, архитектура.

Перцептрон состоял из трех слоев «нейронов» (далее кавычки использоваться не будут, но следует понимать, что речь не идет о биологических нейронах). Первый слой назывался сетчаткой. Нейроны этого слоя назывались сенсорными и моделировали действие рецепторов: их активность определялась не другими нейронами, а входным изображением (один нейрон соответствовал одному пикселю изображения; изображение бралось бинарным: каждая точка изображения была либо белой, либо черной).

Второй слой назывался ассоциативным. Каждый нейрон этого слоя получал сигналы от всех нейронов первого слоя. Эти сигналы домножались на соответствующие веса связей и суммировались (сигнал от некоторого рецептора присутствовал, если соответствующая точка изображения была белой). Если суммарный сигнал превосходил заданный порог, то ассоциативный нейрон активировался и передавал сигнал на нейроны выходного слоя. В простом варианте веса брались равными 1 (возбуждающая связь), −1 (тормозящая связь) или 0 (отсутствие связи). В таком случае ассоциативные нейроны полностью повторяли модель МакКаллока — Питтса и активировались, когда получали сигнал от некоторого минимального числа рецепторов. В еще более простом варианте тормозящие связи к ассоциативным нейронам могли отсутствовать.

Исходно в выходном слое перцептрона присутствовал один нейрон, который суммировал входы от ассоциативных нейронов, помноженные на веса связей (здесь веса уже обязательно брались вещественными). Выходной нейрон активировался, если суммарный сигнал превышал некоторый порог. Подразумевалось, что выходной нейрон должен активироваться, только когда на вход перцептрона подается изображение объекта некоторого класса (в качестве таких объектов выступали буквы английского алфавита). Схема перцептрона Розенблатта приведена на рисунке.

Чтобы перцептрон стал распознавать некоторую букву, нужно установить правильные веса связей, т. е. произвести его обучение. Связи должны постепенно настраиваться при предъявлении изображений как искомого, так и ложных объектов. Здесь могут использоваться модификации правила Хебба: если выходной нейрон активировался при предъявлении ложного объекта, необходимо *уменьшить* веса тех связей, по которым поступал сигнал; если выходной нейрон не активировался при предъявлении искомого объекта, нужно *увеличить* веса тех связей, по которым поступал сигнал. Исходно настраивались связи только между ассоциативными и выходными нейронами. В дальнейшем эти правила нашли более строгое выражение в *методе коррекции ошибки* (но подумайте, не настораживает ли вас что-то в самих этих правилах?) При этом начальные веса связей могут быть заданы случайным образом.

Достаточно очевидно, что при предъявлении одного и того же изображения некоторой буквы перцептрон научится ее распознавать, ведь большие веса будут у тех связей, через

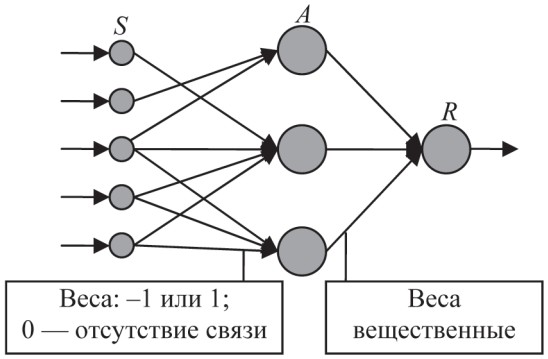


Схема перцептрона Розенблатта

которые распространяется активность при предъявлении искомого объекта, что обеспечивает активацию выходного нейрона. И действительно, перцептрон, в 1957 году смоделированный на компьютере IBM 704, а затем реализованный в 1960 году в виде компьютера MARK 1, показал некоторые способности к обучению и распознаванию. Но всегда ли он сможет научиться отличать искомый объект от ложного? И что будет, если изображение искомого объекта будет варьироваться?

Марвин Минский и Сеймур Паперт в 1969 году выпустили целую книгу «Перцептроны», посвященную математическому анализу этой модели ИНС. В своей книге они пришли к неутешительному выводу: способности перцептронов весьма ограничены. В частности, если входы перцептрона рассматривать как логические переменные, то можно доказать, что перцептрон не всегда сможет выучить логические формулы определенного вида.

Эту проблему часто иллюстрируют на примере задачи об «исключающем или». В этой задаче нужно научиться распознавать «изображения», содержащие всего два пикселя, в зависимости от «яркостей» которых изображения относятся к двум классам: (0, 1) и (1, 0) — истинные изображения, а (0, 0) и (1, 1) — ложные изображения. Соответственно необходимо построить «перцептрон», входной слой которого состоит из двух нейронов, а выходной нейрон оказывается активным или пассивным в зависимости от принадлежности изображения указанным классам (уже здесь возникает вопрос, куда пропал ассоциативный слой, но мы его немного отложим). Проще говоря, необходимо построить перцептрон, реализующий логическую операцию «исключающее или».

Если входные нейроны соединить с выходным нейроном связями с некоторыми весами, то на его вход будет поступать линейная комбинация вида $w_1x_1+w_2x_2$, где w_i — это веса связей, а x_i — сигналы от входных нейронов. Выходной нейрон активируется, когда $w_1x_1+w_2x_2>e$, где e — некоторый порог (активационная функция нейрона является «ступенькой»: на выходе ноль, когда входной сигнал меньше порога, и единица, когда больше). Несложно видеть, что плоскость X_1OX_2 разделяется на две части прямой линией, с одной

стороны от которой находятся объекты, активирующие нейрон, а с другой стороны — не активирующие. Понятно, что нет такой прямой, по разные стороны от которой оказались бы пары точек (0, 0), (1, 1) и (0, 1), (1, 0). Иными словами, требуемого перцептрона не существует.

Но странно, ведь еще МаКаллок и Питтс доказали возможность построения любых логических функций на основе формальных нейронов. Кроме того, и Розенблатт доказал теорему о сходимости перцептрона, согласно которой перцептрон может за конечное число шагов обучиться распознавать любые различимые стимулы, причем обучение будет успешным вне зависимости от начальных весов связей. А ведь еще в 1957 году А. Н. Колмогоров доказал теорему об аппроксимации функций, как следствие из которой можно заключить, что любую непрерывную функцию можно со сколь угодно высокой точностью представить в форме некоторого перцептрона (при некоторых ограничениях на вид активационной функции).

Как можно математически строго доказать противоположные утверждения? В чем же тут дело? А дело в деталях формулировок теорем: в том, накладываются ли какие-нибудь ограничения на структуру сети или нет (подчеркнем еще раз, что и сам перцептрон — лишь весьма частный случай сети на основе простейших формальных нейронов). Например, проблема «исключающего или» не решается одним выходным нейроном, соединенным непосредственно с рецепторами, но вполне решается перцептроном Розенблатта со слоем ассоциативных нейронов. Иногда сеть без ассоциативных нейронов, а только с сенсорными нейронами и одним выходным нейроном ошибочно называют перцептроном, видимо, в связи с тем, что в оригинальном перцептроне связи сенсорных и ассоциативных нейронов были необучающимися. Однако в настоящем перцептроне достаточно всего трех ассоциативных нейронов (или даже двух, если использовать разные пороги активации) помимо двух входных и одного выходного, чтобы реализовать «исключающее или» — достаточно интересно попробовать построить такую сеть самостоятельно, но при этом придется правильно подобрать веса связей между первыми слоями.

Некоторые другие проблемы не будут иметь решения с одним слоем ассоциативных нейронов, если ограничить их количество или, например, запретить связываться каждому из сенсорных нейронов со всеми ассоциативными нейронами сразу. Здесь стоит отметить, что если бы нейроны просто суммировали (с учетом весов связей) входные сигналы и посылали бы результат на выход (т. е. обладали бы линейной активационной функцией), то добавление дополнительных ассоциативных слоев в перцептрон ничего не изменило. Реакция выходного нейрона была бы всегда линейной комбинацией сигналов, поступающих на рецепторы. Добавление дополнительных ассоциативных слоев в этом случае не позволило бы решить проблему «исключающего или». Нелинейность активационной функции здесь имеет принципиальное значение. Благодаря ей увеличение промежуточных слоев позволяет усложнять форму линии, разделяющей классы искомых и ложных объектов. Часто ограничения, присущие однослойным сетям или же сетям с линейными активационными функциями, ошибочно приписывали не только перцептронам, но и произвольным ИНС вообще. Однако результаты Минского и Паперта вовсе не относились только к двухслойным или линейным перцептронам.

Эта, отчасти действительная, отчасти мнимая, ограниченность перцептронов многими была воспринята как приговор искусственным нейронным сетям как таковым. Чрезмерные ожидания сменились столь же необоснованным разочарованием, и массовый интерес к ИНС на время угас. Говорят даже, что критика Минского развернула направление исследований в области ИИ в сторону *символьных вычислений* от направления, называемого сейчас *коннективизмом* (или коннекционизмом, от англ. *connect* — соединять, связывать), по поводу чего сам Минский впоследствии высказывал некоторое сожаление. В рамках коннективизма системы ИИ строятся как совокупности связанных между собой относительно простых элементов. К символьным вычислениям относятся, в частности, экспертные системы, методы эвристического программирования, математическая логика, да и почти вся математика.

Лишь в 1986 году был предложен алгоритм обучения многослойного перцептрона, что стало одной из причин новой

волны интереса к ИНС вообще и к этому типу сетей в частности (конечно, не стоит думать, что в период 1969–1986 гг. исследований в области ИНС не проводилось, тем более что появлялись все новые нейрофизиологические данные). Идею многослойных перцептронов (или, как сейчас говорят, *сетей прямого распространения сигнала*) высказывал еще сам Розенблатт, но без алгоритма их обучения. Сложность обучения многослойной сети состоит в том, что мы знаем правильное значение лишь для выходного нейрона, но не для нейронов промежуточных слоев, поэтому простые правила обучения не работают. Широко используемый алгоритм обучения многослойных перцептронов состоит в обратном распространении ошибки: ошибка на выходе сети как бы распространяется обратно для установления желаемых значений активности в нейронах промежуточных слоев, что позволяет провести коррекцию весов связей. Алгоритм коррекции уже не такой естественный, как в случае с перцептроном с одним ассоциативным слоем, и включает ряд дифференциальных уравнений. Этот алгоритм обучения уже никак не связан с биологией и выводился из сугубо математических соображений.

Однако, как выяснилось, увеличение числа скрытых слоев (при фиксированном числе нейронов) несущественно повышает возможности перцептрона, но заметно усложняет его обучение. И главное, с его помощью остаются нерешаемыми именно те задачи, для которых перцептроны исходно и предназначались. Это задачи распознавания объектов по изображениям. Даже такая простая вещь, как распознавание букв при их произвольном положении на сетчатке, для перцептрона затруднительна. Перцептрон пытается запомнить изображения по-разному смещенной буквы сравнительно независимо. И если ему предъявить изображение этой же буквы, но в том месте на сетчатке, где она раньше не появлялась, перцептрон может ее и не узнать. Формально перцептрон может научиться распознавать все, что угодно. Но в худшем случае ему потребуется столько ассоциативных нейронов, сколько возможно вариантов стимула — это как раз и доказали Минский и Паперт. Так, для всех возможных изображений 1000×1000 пикселей потребуется перцептрон

с $2^{1\,000\,000}$ нейронами (единица с тремястами тысячами нулей), что уже можно сравнить с числом атомов во Вселенной, так как это количество элементов сети, а не их комбинаций. Конечно, на практике не нужно уметь распознавать все произвольные изображения. Кроме того, некоторые изменения стимула, не препятствующие его распознаванию, допускаются.

Также часто говорят, что перцептрону нужно запомнить не все возможные варианты стимулов (и тем более не все потенциально возможные варианты входа — большинство из них могут вообще никогда не встречаться). Вместо этого ему достаточно запомнить просто все встретившиеся примеры изображений: так, за всю свою жизнь человек успевает увидеть столько изображений, сколько вполне уместится на современные запоминающие устройства (и тем более в сам мозг человека). Некоторому перцептрону вовсе не нужно запоминать больше изображений, чем ему в принципе может быть предъявлено, поэтому и чудовищного количества нейронов ему не нужно.

Но дело вовсе не в количестве нейронов — это лишь внешнее проявление того, что перцептрон плохо умеет *обобщать*. Об этом основном недостатке перцептронов говорил еще и сам Розенблатт. Да и Минский в своей критике говорил то же самое, но другими словами. Он это называл проблемой *инвариантов*. Инвариант — весьма продуктивное понятие в математике, означающее некую характеристику объекта, не меняющуюся при заданном типе преобразований этого объекта. К примеру, площадь объекта является инвариантом по отношению к его перемещениям и вращению (но не к изменению масштаба). Итак, перцептрон оказывается неспособным обобщить примеры стимулов, подверженных преобразованиям некоторых типов, чтобы потом уметь распознавать эти стимулы при любых других преобразованиях этих же типов. В таких случаях он может лишь «вызубрить» все варианты, что, как мы знаем, является самым слабым способом обучения. Странно, что критика Минского, в которой просто более строго формулировались те недостатки перцептронов, на которые указывал ранее и Розенблатт, вызвала такую бурную реакцию. К сожалению, сам Розенблатт,

погибший в результате несчастного случая в 1971 году, уже не мог вести дальше эти исследования.

Сейчас многослойные перцептроны широко используются в задачах распознавания, но если речь идет о распознавании изображений, то сначала вручную создается такой алгоритм, на выходе которого строится описание изображения, не меняющееся при его сдвиге, вращении или масштабировании. Если эти описания окажутся удачными, перцептрон сможет научиться распознавать соответствующие объекты. Правда, изображения подвержены не только таким простым, но и многим другим, более сложным преобразованиям, инварианты к которым вручную оказывается построить гораздо труднее.

Итак, известные на сегодня алгоритмы обучения перцептрона (некоторые из которых весьма изощренные) весьма ограничены в своих возможностях и явно не предоставляют общего решения проблемы машинного обучения. Может, недостатки перцептронов связаны с ограниченностью их архитектуры?

МОДЕЛИ АССОЦИАТИВНОЙ ПАМЯТИ

Какое основное ограничение в структуре перцептрона? Число слоев и нейронов в ней не ограничено. Но вот распространение сигнала всегда идет в одном направлении. В результате сеть работает по принципу «стимул—реакция», а ее обучение в чем-то похоже на модель формирования рефлекторной дуги. Почему бы ни сделать сеть с более сложными связями? Конечно, какие-то нейроны в сети должны быть входными, а какие-то — выходными, иначе сеть нельзя будет поместить в некоторую среду или использовать при решении задач. Но все остальные связи между «внутренними» нейронами можно было бы сделать произвольными.

К примеру, если взять за основу перцептрон с двумя слоями ассоциативных нейронов, то можно было бы ввести не только связи от первого слоя ко второму, но и обратные связи от второго слоя к первому, а также связи между нейронами внутри каждого из слоев.

Как же будет вести себя такая сеть? В случае прямого распространения сигнала все выглядит достаточно просто: стимул один раз проходит через сеть, преобразуясь и вызывая какой-то отклик на выходе. Но при наличии обратных связей активация нейронов второго слоя вызовет обратное распространение сигнала к нейронам первого слоя, активность которых в результате этого изменится, они станут посылать новые сигналы на нейроны второго слоя и так далее. Если же и между нейронами одного и того же слоя есть связи, то ситуация еще больше усложнится. Видно, что активность всех нейронов, в том числе и выходных, все время будет меняться, тогда как в перцептроне (сети прямого распространения) сигнал лишь один раз проходит через сеть, на чем работа этой сети заканчивается, пока не изменится входной сигнал.

Если считать, что обычный перцептрон реализует некоторую функцию — логическую (предикат) или вещественную, то сеть с обратными связями будет воплощать функцию (или систему функций), которая вызывает сама себя, т. е. рекурсивную функцию. Такие сети по аналогии называются *рекуррентными*.

В общем случае рекуррентная сеть представляет собой совокупность произвольно связанных между собой нейронов. Какие-то из нейронов считаются сенсорными, и на них подается входной сигнал. С некоторых же (возможно, со всех) нейронов считывается выходной сигнал. Так мы снова приходим к сети самого общего (разумеется, при заданном типе формальных нейронов) вида и к вопросу о том, как она может учиться или хотя бы что с ее помощью можно делать. Математический анализ произвольной рекуррентной сети позволяет кое-что сказать о ее работе. К примеру, можно установить некоторые ограничения, при которых через конечное время работы сети активность ее нейронов перестанет меняться. В этих случаях говорят, что наступила релаксация сети. Но все же о произвольной сети можно сказать так же немного, как, например, и о произвольном алгоритме. Неудивительно, что было предложено много рекуррентных сетей частного вида, решающих вполне определенные задачи. Рассмотрим несколько примеров.

Так, сеть, предложенная Хопфилдом в конце 1970-х годов, представляла собой рекуррентную нейронную сеть, в которой каждый нейрон связан с каждым. Начальная активность нейронов устанавливается в соответствии с некоторым стимулом или образом, по размеру которого и выбирается число нейронов. Отсутствие дополнительных нейронов является очень существенным ее отличием от произвольной рекуррентной сети. Требуется, чтобы наступала релаксация сети, после которой активность нейронов отвечала бы тому же самому входному образу. После настройки весов связей на вход нейронной сети могут подаваться такие же образы, но в зашумленном или поврежденном виде. При этом требуется, чтобы сеть релаксировала в состояние, соответствующее точному оригиналу. Иными словами, сеть должна запомнить образ и уметь его восстановить по зашумленному виду. При этом от сети требуется запомнить одновременно несколько разных образов. Эта задача похожа на задачу распознавания, но немного отличается от нее (на выходе сети должен быть восстановленный образ, а не просто указание на то, к какому классу он относится).

Как, имея набор образов, которые бы хотелось не просто «запомнить», записав в память компьютера, а воспроизвести по зашумленному виду, не перепутав с другими образами, научить это делать рекуррентную нейронную сеть? На первый взгляд, задача может показаться трудной из-за сложной динамики рекуррентной сети. Однако с математической точки зрения она достаточно проста (и требует лишь привлечения операций с матрицами). Основная идея заключается в том, чтобы посмотреть, какими должны быть веса связей, чтобы активность нейронов, отвечающая точным (запомненным) образам, соответствовала состоянию релаксации сети. В результате можно получить формулы для непосредственного вычисления оптимальных весов связей. Сеть может быть сконструирована сразу, без пошагового обучения, как это было в случае перцептрона.

Иногда правило «обучения» сети Хопфилда связывают с правилом обучения Хебба, потому что в состоянии релаксации активными остаются нейроны, между которыми установлены положительные связи. Значит, если мы на вход этой

сети будем все время подавать некоторый образ, задающий распределение активности нейронов, то между постоянно активными нейронами будут устанавливаться положительные связи, что и будет гарантировать запоминание и воспроизведение этого образа. Условно говоря, сеть Хопфилда запоминает изображение, устанавливая положительную связь между белыми точками (и соответствующими им активными нейронами) и отрицательную связь от белых точек к черным точкам («выключенным» в данном образе нейронам). Однако использование правила Хебба с последовательным многократным «предъявлением» образов менее эффективно, чем непосредственное вычисление оптимальных связей по всей совокупности образов.

К сожалению, эта сеть, как и перцептрон, не может воспроизводить образы инвариантно к каким-либо их преобразованиям (например, при заметных сдвигах или поворотах изображения): ведь связи устанавливаются между конкретными «пикселями» сетчатки, а при перемещении образа по сетчатке соответствующие точки образа оказываются в совершенно иных точках сетчатки. Кроме того, «память» этой сети весьма ограничена — она может воспроизводить меньше образов, чем в ней нейронов (а число нейронов в ней ограничено размером самого образа). При этом у данной сети имеется иногда проявляющаяся особенность, заключающаяся в генерации «ложных образов» — состояний релаксации сети, не соответствующих ни одному из образов, предложенных для запоминания.

Интересна возможность использования сети Хопфилда в задачах оптимизации — при поиске приближенных решений NP-полных задач. Об этом типе задач мы много говорили ранее при обсуждении методов эвристического программирования. Использовать данные сети для игры в шахматы затруднительно, но некоторые более простые и в то же время полезные задачи с их помощью решать удается. Итеративная релаксация сети соответствует поиску решения. Это весьма интересно, поскольку показывает возможности применения ИНС и в задачах, традиционно решавшихся в рамках символьных вычислений. Однако обсуждение задач поиска и оптимизации уведет нас в сторону от проблемы обучения,

тем более что при решении этих задач нейронные сети не обучаются, а связи в них жестко выбираются на основе исходных данных задачи.

Существуют и другие архитектуры рекуррентных сетей, реализующие иные модели памяти. К примеру, сеть Коско (описанная в работах 1987 года) состоит из двух слоев с прямыми и обратными связями и реализует *гетероассоциативную память*, восстанавливающую один образ по другому, возможно, никак не связанному с ним по содержанию. При этом входные образы после обучения могут также подаваться в зашумленном виде. В этом смысле сеть Хопфилда *автоассоциативна*, поскольку ассоциирует образ как бы с самим собой, а не с другим образом.

Сеть Коско, как и сеть Хопфилда, хорошо справляется лишь с зашумленными, но не трансформированными образами. Хотя это снижает практическую ценность данных сетей, они весьма интересны как модели памяти.

Из-за широкого использования компьютеров мы привыкли называть памятью различные устройства хранения информации и считать, что человеческая память работает аналогичным образом. Такая *компьютерная метафора* распространена не только в быту; еще раньше она возникла (и до сих пор нередко используется) в когнитивной психологии, занимающейся изучением познавательных (когнитивных) способностей человека. Конечно, на «аппаратном» уровне память человека и компьютера устроена совершенно по-разному, но функцию она выполняет как будто одну и ту же — сохранение информации. И даже традиционное для когнитивной психологии разделение памяти на кратковременную и долговременную сходно с разделением в компьютерах оперативной памяти и носителей информации. Такое разделение иногда критикуется психологами, но оно обосновано нейрофизиологически: кратковременная память не связана с синаптической пластичностью, блокирующейся некоторыми веществами, введение которых мешает долговременному, но не кратковременному запоминанию. Оперативная память, как правило, быстрее, но меньше по объему, и является энергозависимой. Между кратковременной и долговременной памятью есть такие же различия, правда,

объем кратковременной памяти человека делает ее больше похожей на кэш-память (ее иногда называют сверхоперативной памятью).

И все же компьютерная метафора не объясняет особенностей работы памяти человека. Сильнее всего удивляет, почему память человека (в особенности кратковременная) столь несовершенна, почему человек, как компьютер, не может просто записывать и стирать информацию в памяти? Часто говорят, что память человека и компьютера отличается способом записи. Но это не ответ, поскольку тогда возникает вопрос, почему природа использовала такой неэффективный способ записи?

Искусственные нейронные сети не раскрывают тайн человеческой памяти, но некоторые интересные наблюдения позволяют сделать. К примеру, в нейросетевых моделях ассоциативной памяти образы хранятся *распределенно*: каждая связь в такой сети содержит крупицу информации обо всем образе (и обо всех запомненных образах) сразу. Если бы образы покомпонентно хранились в последовательных ячейках памяти, стирание одной ячейки приводило бы к повреждению одной конкретной компоненты одного образа, а зашумление некоторого количества подряд идущих ячеек приведет к полной потере одного из образов. В случае же с ИНС все образы сохранятся, хотя и в немного искаженном виде. Это чем-то похоже на голограмму, даже с помощью небольшого кусочка которой можно воспроизвести записанное изображение целиком, хотя и заметно менее четкое. Аналогию между человеческой памятью и голограммой высказывал К. Прибрам еще в начале 1970-х годов в книге «Языки мозга».

Эта аналогия отчасти объясняет, почему нейрофизиологам так трудно оказывается находить так называемые *энграммы*, или следы памяти, связанные с конкретными событиями или образами. Не стоит, однако, представлять себе мозг как одну большую голограмму, поскольку заметная локализация информации в нем также имеется.

Распределенное хранение информации в нейронных сетях весьма примечательно, но не оно является самым большим отличием памяти человека. В конце концов, в ячейках памяти компьютера тоже можно хранить значения весов связей

некоторой ИНС, что приведет к распределенному хранению информации (и можно вообразить такие ИНС, в которых информация о каждом из образов будет локализована). Так что различия здесь, скорее, в способе представления информации, а не в механизмах памяти.

Другой интересной особенностью нейросетевых моделей ассоциативной памяти является их способность «самостоятельно» делать ошибки. Что же в этом хорошего? Представьте, что испытуемым предъявляют какие-то объекты, которые им нужно запомнить, а потом воспроизвести, а нам нужно промоделировать ошибки при воспроизведении. Если бы мы писали компьютерную программу, которая запоминала те же объекты, что и человек, то для того, чтобы добиться ошибок воспроизведения от этой программы, нам бы пришлось как-то искусственно эти ошибки вводить (на аппаратные сбои памяти при этом вряд ли стоит надеяться). Сети Хопфилда или Коско совершают ошибки в основном из-за своей ограниченной емкости. Если же ограничение на объем памяти ввести в обычной компьютерной программе, она будет запоминать ограниченное число объектов, но не ошибаться при воспроизведении так же, как человек. Однако эта особенность — тоже следствие различий между распределенными и локальными представлениями. В локальных представлениях при нехватке памяти новый образ может просто записывать поверх старого. В распределенных представлениях образы будут как бы накладываться друг на друга.

Интересно также, что при совершении ошибки рекуррентная ИНС сходится к состоянию релаксации в среднем дольше обычного. Нейронная сеть как бы «сомневается» в своем решении. Сходные сомнения часто испытывает и человек. Возможно, ошибки, совершаемые человеком при запоминании, неплохо описываются моделями с распределенным хранением информации. Но ведь причиной ошибок в нейросетевых моделях является не само распределенное хранение информации, а очень небольшой объем памяти из-за малого числа нейронов. Неужели и у человека причиной ошибок является ограниченный объем памяти? Это представляется весьма сомнительным с учетом миллиардов нейронов, которые могли бы участвовать в организации па-

мяти (в том числе и кратковременной, объем которой считается смехотворно малым — 7±2 элемента).

В чем же тогда дело? Может, мы просто неправильно формулируем предназначение человеческой памяти, и оно заключается вовсе не в том, чтобы просто сохранять информацию? На первый взгляд, это предположение кажется нелепым. Но посмотрите: даже простейшие модели памяти на основе ИНС не просто хранят какую-то информацию, а выдают ее в ответ на предъявляемые образы, которые могут быть зашумленными или неполными. Как уже отмечалось, работа такой сети (воспроизведение образа) во многом похожа на распознавание. Но тогда запоминание образа (установление связей в сети) должно иметь много общего с обучением. Мы уже говорили о том, что обучение должно включать некоторое накопление информации (но переработанной, обобщенной), и обучение часто путается с простым запоминанием. Возможно, «запись» информации в память у человека гораздо теснее связана с процессами восприятия и обучения, чем может показаться из компьютерной метафоры. Но для понимания этой связи нужно разобраться, что собой представляет само обучение и восприятие.

ОТ НЕЙРОНОВ ДО ЗРЕНИЯ

С тех пор, как стало известно, что глаз представляет собой оптическую систему, которая строит изображение, философы задавались следующим вопросом: находится ли в голове некий человечек, гомункулус, который смотрит на это изображение. И если он на него смотрит, то, в свою очередь, находится ли у него в голове еще один человечек, который также смотрит на следующее изображение, и так далее. Очевидно, лишь одного того факта, что глазом формируется изображение, совершенно недостаточно для объяснения процесса зрения. Всю глубину непонимания этого процесса показывает следующий вопрос: как и где в мозгу перевернутое изображение, формируемое глазом, переворачивается обратно? Когда-то эта проблема казалась серьезной,

но сейчас с алгоритмической точки зрения понятно, что это не проблема вовсе или, по крайней мере, самая мелкая проблема, о которой нужно думать. Отметим еще раз, насколько более продуктивными математическая теория алгоритмов сделала наши рассуждения о мышлении.

Даже тот факт, что человек, надевший и непрерывно носивший специальные очки, переворачивающие изображение, через некоторое время перестает видеть мир перевернутым, кажется не слишком удивительным. Алгоритмически такую функцию реализовать гораздо проще (и она уже реализуется во многих цифровых камерах), чем распознавание. Примечательным является лишь то, что она оказывается доступной человеческому мозгу, несмотря на ее биологическую бесполезность. То, что мозгу требуется заметное время на переворачивание изображения, видимо, говорит о том, что эта функция не заложена в него заранее, но формируется только при возникновении такой необходимости (в отличие от камер, где она запрограммирована разработчиком).

Человеку субъективно кажется, что зрительное восприятие — нечто очень простое. Слишком легко оно ему дается. Даже перемножение двух трехзначных чисел кажется сложнее, чем их прочтение с листа бумаги. Мало кто понимал истинную сложность этого процесса до того, как стали проводиться попытки его алгоритмического воспроизведения. Что же происходит после того, как на сетчатке глаза формируется изображение?

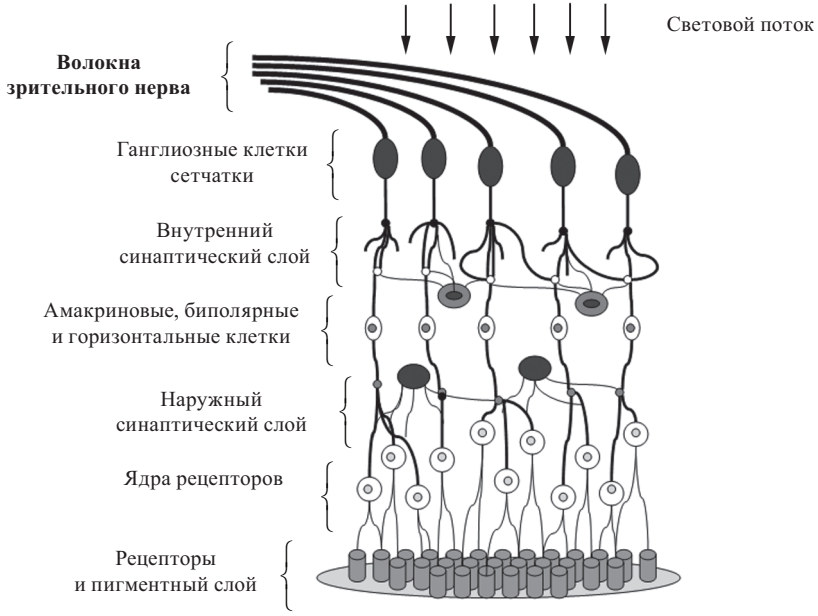
Первая реализованная искусственная нейронная сеть — перцептрон — предназначалась для моделирования сетчатки, которая содержит не просто фоточувствительные рецепторы — палочки и колбочки, — но также и несколько слоев разнотипных нейронов. Такая слоистость была известна и во времена изобретения перцептрона, но на этом его сходство с реальной сетчаткой заканчивается. Может быть, ограниченность способности обучаться, относящаяся ко всем рассмотренным типам ИНС, вызвана тем, что в них заложены ранние сведения о функционировании естественных нейронных сетей? Неужели более чем за полвека не появилось новых сведений в нейробиологии? Конечно, сведений появилось много.

К примеру, достаточно быстро стало известно, что в действительности нейроны в сетчатке не обучаются распознавать разные объекты. Они выполняют лишь самые первые шаги обработки изображений и посылают предобработанное изображение для гораздо более детального анализа по зрительному нерву в обширную зрительную кору головного мозга, занимающую весьма заметную долю общей площади коры мозга — 15 %.

Интересно, что в глазу человека насчитывают свыше 100 млн рецепторов, что условно соответствует регистрации изображения 10000×10000 пикселей. В то же время зрительный нерв состоит из аксонов всего лишь миллиона нервных клеток (называемых *ганглиозными*). Зачем иметь столько рецепторов в сетчатке, чтобы передавать в мозг меньше 1% объема регистрируемой информации (если, конечно, по миллиону нервных волокон не успевает каким-то хитрым образом передаваться вся информация, регистрируемая ста миллионами клеток)?

Природа обычно экономна, и вряд ли в глазу было бы так много рецепторов, если бы они на самом деле не использовались. Хорошо известно, что изображения обладают большой информационной избыточностью. Упрощенно говоря, это выражается в том, что цвета соседних точек на изображениях обычно отличаются мало. Исключения составляют точки на границах объектов, где происходит резкое изменение яркости или цвета. Если бы мы вычли из яркости каждой точки среднюю яркость в ее окрестности, то получили бы новое «изображение», на котором были бы отображены контуры объектов, а все остальные точки стали темными. Изображения, полученные в соседние моменты времени, часто похожи, и «соседние кадры» можно тоже вычитать друг из друга для передачи в мозг их различий.

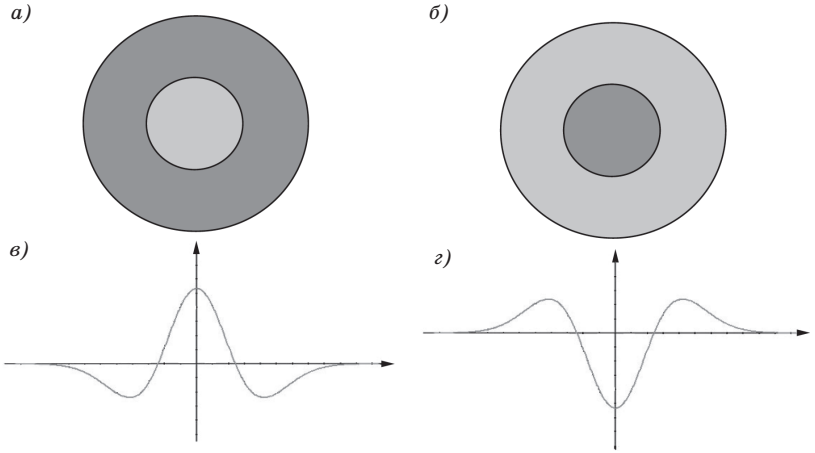
Структура связей между нейронами сетчатки хорошо изучена (в очень приближенном виде она представлена ниже на рисунке). Интересно, что слой фоторецепторов расположен наиболее далеко, а слой клеток, аксоны которых формируют зрительный нерв (уходящий через слепое пятно обратно в мозг), наиболее близко к поверхности глаза. Также изучены *рецептивные поля* отдельных нейронов, т. е. те области сетчатки, с которых в данный нейрон попадает информация.



Структура сетчатки

Можно было бы обсудить работу всех слоев сетчатки, но особый интерес представляют рецептивные поля ганглиозных клеток, посылающих сигналы по зрительному нерву в мозг, поскольку с их помощью можно понять функции всей сетчатки по переработке зрительной информации. Нейрофизиологами изучались отклики ганглиозных клеток в ответ на различные визуальные стимулы. В частности, Куффлер подключал к разным ганглиозным клеткам электрод и, проецируя в разные места сетчатки маленькое пятнышко света, отыскивал на сетчатке области, освещение которых увеличивало или подавляло активность конкретной ганглиозной клетки.

Выяснилось, что рецептивные поля многих ганглиозных клеток круглые и состоят из центра и периферии (как показано на рисунке), причем эти клетки бывают двух типов: у клеток с так называемой on-реакцией засветка центра рецептивного поля приводит к повышению активности клетки, а засветка периферии — к подавлению ее активности, в то время как у клеток с off-реакцией все в точности наоборот.



Структура рецептивных полей ганглиозных клеток (темным цветом показаны области торможения, светлым — области возбуждения): *а* — случай on-реакции; *б* — случай off-реакции; *в*, *г* — сечения карт рецептивных полей в случае on- и off-реакций соответственно

Таким образом, если в центр рецептивного поля клетки с on-реакцией проецировать пятнышко света, то при увеличении размеров этого пятнышка активность клетки будет возрастать, пока пятно не заполнит весь центр. Когда же пятно начнет заполнять периферию, то активность клетки будет уменьшаться. При равномерной засветке всего рецептивного поля ганглиозная клетка практически не будет проявлять активности.

Видно, что ганглиозными клетками действительно вычисляются усредненные разности яркостей точек и их соседей. Кроме того, имеется еще один тип ганглиозных клеток, которые вычисляют не пространственные, а временны́е изменения яркости, и возбуждаются только при наличии движения на изображении в соответствующей области сетчатки.

Конечно, несмотря даже на это, довольно трудно с уверенностью утверждать, что именно происходит в сетчатке, ведь свойства сетчатки не так просты. К примеру, рецептивные поля ганглиозных клеток сильно варьируются по размеру и другим параметрам, а сами клетки получают не только прямые сигналы от рецепторов, но и обратные из мозга. Однако гипотеза о том, что нейронами сетчатки выполняются операции по уменьшению избыточности (которые часто

представляются как пространственное и временно́е дифференцирование изображений), является весьма популярной.

Зрительная информация от нейронов сетчатки через некоторые дополнительные структуры (такие, как наружное коленчатое тело) попадает в зрительную кору, которая разбивается на большое число зон. Наиболее изученной сейчас является *первичная зрительная* (или *стриарная*) кора. Это участок зрительной коры, нейроны которого первыми в коре мозга получают зрительную информацию. Интересно, что в стриарной коре порядка 200 млн клеток, которые принимают информацию всего от пары миллионов ганглиозных клеток сетчатки двух глаз. Чем же занимается такое количество нейронов? Установлено, что значительная часть этих нейронов (около 70 %) откликаются на полосы и линии, причем каждый нейрон активируется, только если такие линии появляются в определенной области на сетчатке и обладают определенной ориентацией (основные свойства этих нейронов были открыты в 1960-х годах Д. Хьюбелом и Т. Визелом, за что им в 1981 году была присуждена Нобелевская премия). На естественных изображениях максимальный отклик возникает на протяженных границах объектов. Если считать, что ганглиозные клетки просто выделяют точки на границах объектов (точки, в которых нарушается пространственная однородность изображения), то работа клеток стриарной коры выглядит вполне уместной в качестве следующего шага обработки. Можно сказать, что эти клетки описывают изображения в терминах некоторых простейших структурных элементов, таких как отрезки прямых линий. Полагается, что в последующих зонах зрительной коры нейроны могут реагировать на более сложные формы, например треугольники. А более высокие зоны строят описания объектов в терминах еще более сложных элементов.

Конечно, работа зрительной коры гораздо сложнее: в ней происходит анализ и цвета, и текстуры, и движения; информация от двух глаз объединяется для осуществления стереозрения; работает множество других механизмов. Высшие зрительные функции все еще остаются загадкой. И самое главное, до сих пор не известно, как происходит обучение. До конца неясно даже, что в структуре зрительной системы за-

ложено генетически, а что формируется под влиянием опыта. Хотя структура зрительной системы у человека продолжает формироваться до 4–5 лет, это может быть как реализацией генетической программы, лишь немного адаптирующейся к окружению, так и детальным обучением, в результате которого создаются основные связи зрительного тракта.

Существует много разных экспериментов, которые проводились с целью ответить на этот вопрос. Классическим считается эксперимент, выполненный в 1970 году К. Блейкмором и Дж. Ф. Купером. В этом эксперименте котят содержали в окружении, обеспечивающем восприятие только чередующихся черных и белых полос. Как оказалось, у котят потом в первичной зрительной коре наблюдались лишь клетки, чувствительные к вертикальным полоскам, а клеток, чувствительных к элементам другой ориентации, обнаруживалось гораздо меньше, чем в норме. В других экспериментах вместо изменения окружения использовались специальные очки, причем один глаз видел преимущественно горизонтальные, а другой — вертикальные контуры. В результате клетки, получающие разную информацию от разных глаз, оказались более чувствительными к контурам своей ориентации. Интересно, что эти изменения могут быть заметны не только по функционированию нейронов, но и визуально, под микроскопом. Хьюбел отмечает поразительность того факта, что возможно вызывать заметные физиологические и морфологические изменения в нервной системе лишь с помощью информационного воздействия без реального физического вмешательства. Это следует признать верным, по крайней мере, для первых лет жизни.

Но даже эти (равно, как и некоторые другие) эксперименты не позволяют судить о том, насколько пластичной является зрительная система в процессе своего формирования в ранний постнатальный период. Многие исследователи полагают упомянутые эксперименты убедительным свидетельством большой пластичности даже первичной зрительной коры, в результате которой формируются связи, не заложенные генетически. Не менее правдоподобна и другая гипотеза, согласно которой депривация зрительной системы (т. е. лишение ее всего богатства естественных стимулов)

ведет лишь к деградации нейронов, в норме реагирующих на те стимулы, которые отсутствуют в обедненной среде, и перераспределению ресурсов между работающими нейронами. Здесь полагается, что бедная среда ведет к исчезновению полезных связей, заложенных генетически. При этом сомнительной считается возможность создания такой обогащенной среды, в которой в первичной зрительной коре будут появляться нейроны, реагирующие на более разнообразные, чем в норме, стимулы. Бесспорно, для более высоких зон зрительной коры влияние опыта должно повышаться.

С другой стороны, некоторые эксперименты показывают, что если после рождения животному зрительный нерв «подсоединить» к слуховой коре вместо зрительной, то оно научится видеть с помощью слуховой коры. Остается пока неясным, связано ли это с активацией генетических программ, содержащих сведения об алгоритмах обработки информации соответствующей сенсорной модальности, или же с чрезвычайной обучаемостью нейронов коры.

Несмотря на неоднозначность интерпретации, эксперименты Блейкмора и Купера вместе с другими сведениями о работе первичной зрительной коры, в частности, показанная Хьюбелом и Визелом чувствительность этих нейронов к ориентациям линий и краев, привели к развитию моделей распознавания образов на основе обучающихся искусственных нейронных сетей. Именно на эти эксперименты ссылается Кунихико Фукусима в статье 1975 года, в которой он предложил нейронную сеть под названием «Когнитрон».

Когнитрон имеет определенные сходства с перцептроном. Это тоже сеть прямого распространения, состоящая из нескольких слоев. Однако правила ее обучения другие. Во-первых, когнитрон учится без учителя (или самообучается), т. е. ему даются только сами изображения и не сообщается, совершил ли он ошибку в результате своей работы. В связи с этим когнитрон не осуществляет коррекцию связей на основе информации о своих ошибках, но организует свои связи так, чтобы сходные образы классифицировать одинаково. Конечно, когнитрон является не единственной сетью, обучающейся без учителя. Таковыми являются, например, самоорганизующиеся карты Кохонена, веса в которых так-

же настраиваются так, чтобы сеть одинаково реагировала на похожие образы.

Во-вторых, в процессе обучения связи усиливаются не просто между парой нейронов, активирующихся одновременно, но между парой нейронов, активность которых максимальна в некоторой пространственной области (обучение работает по принципу «победитель забирает все»). Фукусима обосновывает такой способ обучения с точки зрения работы глиальных клеток, заполняющих пространство между нейронами (хотя сами эти клетки и не вводятся в модель в явном виде): глиальные клетки питают в своей окрестности наиболее активную пару нейронов, через связь между которыми проходит сигнал. Это отличается от простейшего варианта правила Хебба, в котором усиливается существующая связь между любыми одновременно активными нейронами. Такая модификация правила обучения приводит к тому, что каждый нейрон начинает реагировать на определенный элемент в определенной области изображения (например, на линию определенной ориентации) после многократного предъявления одного и того же стимула. Нейроны последующих уровней реагируют на комбинации простейших стимулов, расположенных определенным образом. При этом в отличие от перцептрона в процессе обучения когнитрона без особых трудностей настраиваются связи между нейронами всех уровней.

Интересно, что если когнитрону предъявлять только стимулы, состоящие из линий вертикальной ориентации, то после обучения нейроны первого слоя будут реагировать только на вертикальные линии. Чем больше разнообразие предъявляемых стимулов, тем больше будет и разнообразие рецептивных полей нейронов. Бесспорно, все это — и повышение сложности воспринимаемых стимулов с уровнем, и особенности обучения — имеет гораздо большее сходство со свойствами зрительной системы, чем у перцептрона. Можно было бы даже сказать, что когнитрон более гибок, чем реальная первичная зрительная кора, ведь нейроны его первого уровня могут научиться реагировать не только на линии разной ориентации, но и почти на любые другие локальные особенности изображений (правда, только бинарных). Пластичности такого уровня не удается обнаружить даже

на самых ранних этапах развития первичной зрительной коры. Однако когнитрон, как и все ранее рассмотренные ИНС, не способен обучиться распознавать изображения при их смещениях и других преобразованиях, если в обучающей выборке не было изображений, преобразованных почти так же, как и распознаваемое изображение.

В целях преодоления этого фундаментального ограничения Фукусима к 1980 году разработал «Неокогнитрон». В этой ИНС моделировалось еще одно свойство первичной зрительной коры, которое состоит в наличии в ней так называемых *простых* и *сложных* клеток. Различие между этими клетками заключается в том, что простые клетки реагируют на стимул, находящийся только в определенной точке изображения. Если стимул отклоняется от этой точки, то реакция простой клетки на него ослабевает. Сложные же клетки реагируют на свой стимул почти независимо от его положения в некоторой области изображения — рецептивном поле сложной клетки.

В когнитроне сложные клетки используются для обеспечения инвариантности к сдвигам и отчасти — повороту и масштабу. Модели этих клеток собирают информацию с однотипных простых клеток, реагирующих на один и тот же стимул, находящийся в разных местах изображения. Такая модель сложных клеток требует больших объемов вычислений. Сейчас существуют модели, позволяющие воспроизвести свойства сложных клеток без сбора информации с простых клеток, отвечающих всем возможным положениям некоторого стимула. Но не это главное. Однотипность простых клеток приходится задавать заранее. Понятно, что на основе некоторой простой клетки, связи которой настроены в процессе обучения на реакцию на определенный стимул, проблематично построить сложную клетку, так как нет других простых клеток, реагирующих на тот же стимул, но появляющийся в других местах изображения. Даже если когнитрону предъявить сто одинаковых стимулов, но с разными смещениями, он не сможет распознать сто первый стимул с новым смещением, т. е. он не способен выполнить обобщение. Общую структуру связей в неокогнитроне между простыми и сложными клетками приходится задавать заранее. Это

могло бы служить объяснением, почему на нижних уровнях обработки информации в зрительной системе не происходит обучения, в результате которого клетки могут реагировать на разнообразные стимулы, и связи между клетками во многом определяются генетически. Однако в неокогнитроне даже инвариантность к сдвигу достигается не в результате обучения, а задается априорно (заранее). Возможно, в зрительной системе этот тип инвариантности тоже задан генетически (поскольку это важно для выживания), однако нельзя не согласиться, что человеческий разум способен формировать новые, куда более сложные, инварианты. А наделение компьютера именно такими способностями и является целью машинного обучения как научной области.

В этом смысле неокогнитрон тоже не дает решения одной из центральных проблем машинного обучения. Его способность к инвариантному распознаванию за счет специально разработанной структуры связей не слишком интересна, поскольку в компьютерном зрении этот же результат был достигнут гораздо раньше с использованием традиционных (не нейросетевых) алгоритмов. Вообще, сейчас в компьютерном зрении ИНС не очень популярны, поскольку с помощью обычных алгоритмов удается достичь результатов заметно более эффективным образом. В то же время нельзя не признать, что сведения из нейрофизиологии и психофизиологии зрительного восприятия сыграли большую роль в развитии этой области. При этом некоторые известные механизмы зрительного восприятия до сих пор редко воспроизводятся в компьютерном зрении. Одним из таких важных механизмов является обратная связь между уровнями восприятия, которая для воспроизведения требует сетей не прямого распространения, а рекуррентных.

НЕОДНОЗНАЧНОСТЬ И АДАПТИВНЫЙ РЕЗОНАНС

Исследования зрительной системы показывают, что обработка информации в ней организована иерархически, т. е. с разделением по уровням: на начальных уровнях нейроны

реагируют на простые стимулы, и нейроны каждого последующего уровня собирают информацию с групп нейронов предыдущих уровней, избирательно реагируя на все более сложные стимулы. Почему же зрительная система устроена таким образом?

Можно полагать, что такая структура зрительной системы обусловлена иерархичностью организации самого мира. Ведь, действительно, каждый объект состоит из совокупности меньших объектов: так, лес состоит из деревьев, деревья — из ствола и ветвей, которые сами по себе образуют иерархическую структуру, и так далее. Однако даже до выделения каких-либо целостных объектов в зрительной системе присутствует много уровней обработки, связанных с описанием изображения в терминах контуров, простых и составных структурных элементов, а также цветовых и текстурных свойств и т. д. Зачем выделяются на изображении такие элементы? Почему нельзя непосредственно распознавать разные объекты и уже на их основе строить иерархические описания сцен?

То, что объект следует распознавать по совокупности его частей, кажется вполне естественным. Многие начинающие специалисты в автоматическом распознавании изображений высказывают такую идею. Так, лицо можно распознать как совокупность определенным образом расположенных глаз, носа и ушей, дом — как совокупность стен, окон, дверей и крыши, и т. д. Однако эта очевидность обманчива: ведь как распознавать простые элементы? Откуда мы знаем, что некоторый объект именно ножка стола, а не просто палка или ручка от какого-то спортивного снаряда? Да и как установить, что некоторая группа пикселей на изображении — отдельный объект?

Проблема заключается в том, что изображение исходно представляется в зрительной системе почти как в компьютере: каждая палочка или колбочка отвечает лишь за один «пиксель» изображения. Как из этих отдельных пикселей сформировать целостный образ? Можно было бы перебирать все возможные области на изображении и определять, на какой из объектов каждая из таких гипотетических областей похожа. Несложно догадаться, что число разных вариантов

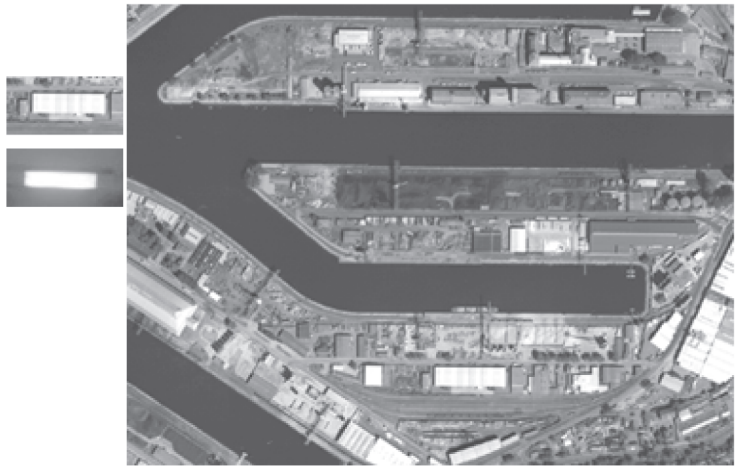
разделения изображения на области чрезвычайно большое — оно экспоненциально растет с ростом размера изображения, поэтому задача интерпретации изображений является NP-полной. Это означает то, что идеальное решение этой задачи невозможно за обозримое время. Но ведь зрительная система как-то с ней справляется! Значит ли это, что мозг все же может решать быстро NP-полные задачи?

Если присмотреться к зрительной системе (равно как и к другим сенсорным системам человека), то следует признать, что она работает очень хорошо, но не идеально: нередко мы не можем сразу понять, что же все-таки видим, а иногда и совсем обманываемся. Если не считать, что зрительная система с помощью какого-то чуда идеально решает NP-полную задачу интерпретации изображений, то стоит задуматься, не связаны ли особенности ее строения (в частности, иерархичность) с эффективным приближенным решением задачи интерпретации...

И можно смело утверждать, что это действительно так: если бы естественные нейронные сети обладали чудесной способностью «неалгоритмического» решения NP-полных задач, зрительной системе не нужно было бы обладать столь сложной структурой. Да и в компьютерном зрении иерархичность используется для той же цели — добиться решения задач анализа изображений за реалистичное время при незначительных потерях в качестве решения. Почему же происходит потеря качества?

По сути, в иерархических методах общая задача анализа изображений (или другой сенсорной информации) разбивается на подзадачи, которые решаются независимо, например, каждый нейрон стриарной коры реагирует на свой локальный стимул без учета всего изображения. Лишь на следующих уровнях анализа реакция на эти стимулы объединяется для того, чтобы выделить более сложные стимулы. Но откуда известно, что простые стимулы без учета окружения будут выделены оптимально?

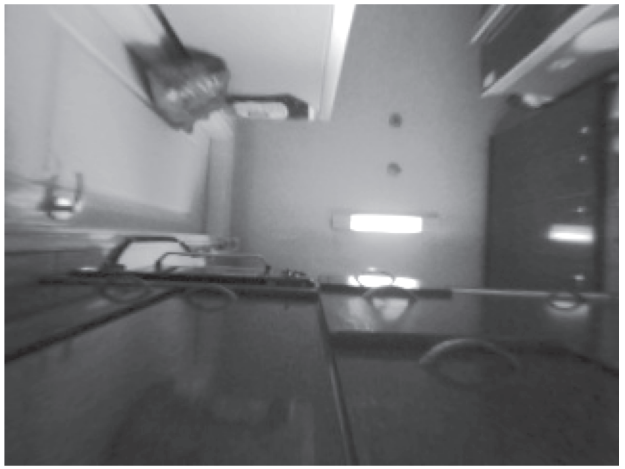
К примеру, хорошо известно, что если взять фрагмент изображения с небольшим объектом, то может оказаться практически невозможно распознать объект, который легко узнается на исходном изображении. Так, на рисунке при-



Распознавание объекта по контексту

ведены отдельно два объекта, один из которых присутствует на полном изображении.

Как только мы видим, что это аэрокосмический снимок, мы можем легко распознать в белом прямоугольнике здание. Второй отдельный объект отсутствует на этом изображении, и узнать его не так просто. Хотя по виду он мало отличается от первого объекта, в действительности он является вовсе не домом. Чтобы распознать его, достаточно взглянуть на фотоснимок потолка.



Снимок потолка с лампой

Получается, что объекты нередко нельзя распознать по отдельности. Скажем, ножку, отделенную от стула, узнать гораздо труднее, чем ножку, присоединенную к стулу. Но как же тогда распознается весь стул, если не как совокупность правильно расположенных ножек, сиденья и спинки? Если на каждом уровне иерархического анализа существует вероятность неправильно распознать какой-то элемент изображения или объект, то с увеличением уровней качество распознавания должно все больше ухудшаться! Если мы просто распознаем маленькие детали и последовательно объединяем их в более масштабные объекты, то почему мы все же ножку в составе стула узнаем лучше, чем отдельную ножку — ведь она должна распознаваться раньше, чем весь стул? Все то же самое относится не только к объектам, но и к промежуточным структурным описаниям: если на зашумленном изображении представлен квадрат, то выделенные по отдельности его границы окажутся непараллельными. Можно будет даже ошибиться с числом его вершин из-за сглаженности углов на контурах.

Из-за этой трудности рядом исследователей была поддержана гипотеза, согласно которой изображение, наоборот, сначала интерпретируется в целом, сюжетно (как лес, город, комната, т. е. некоторый сюжет), а затем уже распознаются детали этого изображения от крупных к маленьким. Эта гипотеза объясняла бы надежность зрительного восприятия, если бы описывала, как именно можно распознавать сцены до распознавания отдельных объектов. Иногда говорят, что разные полушария мозга руководствуются разной стратегией интерпретации сенсорной информации: правое полушарие воспринимает изображение одновременно в целом, а левое последовательно анализирует детали.

Существуют такие повреждения правого полушария мозга, при которых нарушается способность формировать целостный образ из отдельных деталей. Один случай такой агнозии (расстройства в распознавании и восприятии) описывает О. Сакс в упоминавшейся уже книге «Человек, который принял жену за шляпу». Больной был способен распознавать отдельные элементы, признаки изображений, но не видел объекты в целом, из-за чего ему приходилось

выполнять эту операцию на уровне сознания. Например, он был способен «узнавать» лишь трех коллег по работе, используя такие признаки, как наличие у них нервного тика, большой родинки на щеке или по чрезвычайной худобе. В других случаях, при повреждении левого полушария, человек может распознавать рисунок из совокупности случайно расположенных окон, дверей, крыши как дом, но быть не в состоянии сказать, чем этот рисунок отличается от правильно нарисованного дома.

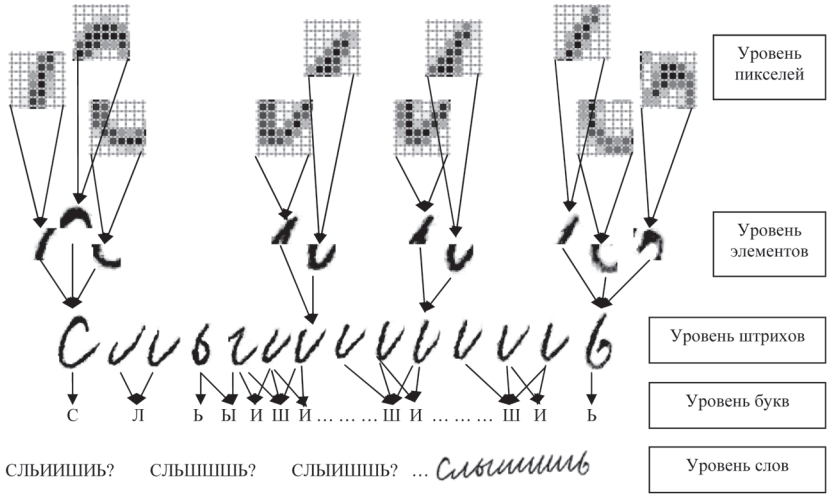
Интересно, что больной, описанный Саксом, будучи художником, при развитии болезни правого полушария постепенно в своей живописи все больше склонялся к кубизму, связанному, видимо, с активностью левого полушария. Напротив, такое направление, как импрессионизм, можно связать с большей активностью правого полушария. Такие примеры подтверждают то, что стратегии обработки информации в разных полушариях могут отличаться. Вообще говоря, многие направления в живописи связаны, видимо, с гиперболизацией тех или иных механизмов зрительного восприятия. Хорошо известно влияние на искусство такого феномена, как *синестезия*, впервые описанная в XIX веке Фрэнсисом Голтоном, двоюродным братом Чарльза Дарвина. Сейчас известно, что синестезией обладало достаточно много музыкантов и художников. При синестезии, по метафоре В. С. Рамачандрана (приводящего об этом феномене много фактов в своей книге «Рождение разума. Загадки нашего сознания»), имеется как бы «перекрест проводов» в мозгу, в результате чего происходит «перетекание сигнала» между областями мозга, занимающимися обработкой информации разных модальностей. Как следствие, человек способен видеть звук или слышать цвет. Крайне интересно и то, что были случаи цветовой синестезии у дальтоников, лишенных части цветовых рецепторов в глазу (это подчеркивает то, что переживание цветовых ощущений происходит в зонах коры, предопределенных генетически).

Несмотря на возможные разные стратегии обработки информации, на сетчатке все же изображение представляется в виде отдельных точек, каждая из которых активирует свою клетку-рецептор. При этом информация от отдельных

рецепторов как-то должна объединяться, «интегрироваться». И действительно, если фоторецепторы реагируют на отдельные точки, то последующие нейроны объединяют информацию от нескольких нейронов предыдущего уровня и реагируют на все более сложные стимулы.

Такое постепенное объединение элементов изображения (или другой сенсорной информации) выглядит весьма правдоподобно, но все же правильное распознавание деталей без распознавания целого выполнять может только весьма ненадежно. Рассмотрим пример, на котором это было бы отчетливо видно. На приведенном ниже рисунке показаны уровни, на которые мог бы делиться анализ некоторого написанного от руки слова. Сначала отдельные пиксели могли бы объединяться в простейшие элементы, которые бы затем объединялись в штрихи, а те, в свою очередь, — в буквы. По буквам было бы уже просто распознать слово.

Даже если считать, что штрихи формируются идеально (хотя на самом деле здесь ошибиться очень просто, анализируя лишь небольшие фрагменты изображения), объединить их в буквы без ошибки очень сложно. Если смотреть в середину приведенного на рисунке слова, действительно невозможно сказать, какие штрихи входят в букву «ш», а какие — в букву «и».



Восходящее распознавание текста

Такая неопределенность возникает на всех уровнях и во всех модальностях. К примеру, в одном из экспериментов, проводимых учеными, испытуемым на слух предъявлялась фраза, в которой некоторый звук в некотором слове был сильно зашумлен. Этот звук нельзя было определить по его собственному звучанию. Более того, для «испорченного» слова существовало несколько вариантов, отличающихся только одной буквой (например, «?орт» можно услышать, как «порт», «торт», «сорт», «корт», «борт», ...). В зависимости от всей фразы «испорченное» слово воспринималось испытуемыми по-разному, даже если во всех случаях оно было одним и тем же. При этом у человека часто и мысли не возникало, что слово можно было распознать по-другому. Интересно, что результаты эксперимента не менялись, даже если зашумлялся первый звук первого слова.

Видно, что простой иерархичности недостаточно и нельзя на каждом уровне обработки делать однозначный выбор. Может быть, с каждого предыдущего уровня на последующий передается не одна, а несколько гипотез для каждого анализируемого фрагмента данных? В предыдущем примере это могут быть все возможные слова, заканчивающиеся на «...орт». На следующем уровне могут строиться разные фразы для всех возможных вариантов этого слова и из них как-то выбирается наиболее осмысленная. Но представим, что в предложении есть два слова, каждое из которых допускает по четыре варианта интерпретации. Тогда возможных вариантов предложений будет 16. Если же таких слов три, то вариантов будет 64. Такое вполне может быть. Ведь каждый из звуков по отдельности часто сложно определить однозначно. А если в некотором слове два или три звука неоднозначны?.. Посмотрим на написание слова «слышишь» от руки. Сколько вариантов последовательностей букв там возможно! Конечно, многие варианты не соответствуют реальным словам. Но можно ли их заранее отсекать? Вдруг во фразе присутствует какое-то новое слово? Видно, что число возможных гипотез будет расти лавинообразно, а локальный контекст на каждом уровне может обеспечить лишь частичное отсечение неправдоподобных интерпретаций.

Контекст, однако, не ограничивается теми данными, которые мы воспринимаем в конкретный текущий момент времени. Обычно у нас есть определенные ожидания, что можем увидеть или услышать, поскольку мы осведомлены о своем окружении. Именно эти ожидания и позволяют устранить неоднозначность в интерпретации сенсорной информации на наиболее высоком уровне восприятия. Наличие таких ожиданий может быть четко установлено, когда выбор способа восприятия неоднозначного стимула происходит не за счет непосредственного контекста, а за счет предварительной установки. Например, в психологии хорошо известен эффект *перцептивной готовности*, при котором начальная установка сильно сказывается на восприятии. Часто этот эффект демонстрируют на примере слов «желтый», «зеленый», «красный», «синий», написанных не теми цветами, которые эти слова обозначают. Интересно, что слова эти прочитать заметно легче, чем назвать цвет, которым они написаны. Конечно, от того, что синим цветом написано слово «красный», нам синий цвет краснее казаться не будет. Этот пример, хоть и весьма эффектный, но не вполне удачный. Зато в жизни встречаются и более подходящие примеры. Всем попадались, скажем, кроссовки с надписью «Abibas», которую вполне можно принять за название известной фирмы. Когда нам такая надпись встречается в другом контексте (например, как здесь) шансов ошибиться с ее прочтением меньше.

Как уже упоминалось, по контексту также разрешается и омонимия слов. Осмысление предложения начинается с отдельных входящих в него слов. Но каждое слово может иметь несколько значений, поэтому нужны какие-то механизмы перебора разных комбинаций значений слов в поисках наиболее осмысленной. Иногда смысл слов может быть непонятен только по одному предложению и нужен более широкий контекст. Так, к примеру, если в некотором тексте речь идет про некоторую девочку, то при прочтении предложения: «У нее была небольшая коса» — у читающего уже будет определенная установка, и при восприятии этого предложения вряд ли возникнет, например, образ девочки, держащей в руках сельскохозяйственное орудие.

Наиболее ярко эффект перцептивной готовности заметен на негативных примерах — когда из-за него наше восприятие ошибается. Однако в большинстве случаев использование предварительных установок, ожидания, общего контекста играет положительную роль, позволяя нам не тонуть в море альтернативных интерпретаций сенсорной информации, не приглядываться каждый раз к надписям или предметам. Все это хорошо подтверждает гипотезу о том, что в процессе восприятия приближенно решается NP-полная задача и при этом активно используется весь имеющийся контекст для сокращения числа гипотез. Эффект перцептивной готовности показывает, насколько языковая установка через оптимизацию процедур обработки сенсорной информации способна оказывать влияние на восприятие; в этом смысле упоминавшаяся гипотеза лингвистической относительности о зависимости нашего восприятия мира от языка не так уж лишена смысла.

Но как все это работает? Если на каждом этапе обработки для каждого кусочка данных порождается некоторое число гипотез, то на последующих этапах обработки при объединении этих кусочков в более крупные образования число гипотез будет расти в геометрической прогрессии и их будет весьма затруднительно сравнивать с имеющимися ожиданиями.

Стефан Гроссберг, занимаясь исследованием особенностей человеческого восприятия, еще в 1970-х годах предложил концепцию восприятия, названную им *адаптивным резонансом*. Суть идеи адаптивного резонанса в том, что гипотезы, активировавшиеся на более низких уровнях, не просто *передаются* на более высокие уровни для дальнейшего анализа, а *усиливают* удовлетворяющие им гипотезы более высоких уровней. В то же время и «активные» гипотезы более высоких уровней усиливают допускающие их гипотезы более низких уровней. Все это достаточно естественно представляется в виде рекуррентной нейронной сети, в которой уровень активности нейронов соотносится с уровнем принятия соответствующих гипотез. Нейрофизиологические данные также подтверждают существенную роль обратных связей, идущих от коры мозга к органам чувств и передающих информацию в направлении, обратном направлению прямой обработки сенсорной информации. Иногда число обратных связей оказывается даже больше, чем прямых.

Нейроны верхних уровней могут активироваться под воздействием активности нейронов нижних уровней. Тогда их роль будет заключаться в постепенной интеграции информации и разрешении неопределенности только на основе текущего контекста. При этом, правда, снятие неопределенности может быть затруднено. Но активность этих нейронов может быть также вызвана имеющимися ожиданиями (которые в восприятии человека должны играть очень большую роль). Если сила ожиданий слишком высока, то в крайнем случае активность нейронов верхнего уровня будет оставаться неизменной, подстраивая под себя через обратные связи активность нейронов всех прочих уровней вне зависимости от того, что происходит в действительности. Это можно связать с работой воображения или с ситуацией, в которой человек видит и слышит лишь то, что хочет или готов воспринять. Еще один феномен, который может быть объяснен распространением активности с верхних уровней вниз, — галлюцинации, часто возникающие при сенсорной депривации (лишении органов чувств входных сигналов). Действительно, если с нижних уровней активность не распространяется, то преобладать будут ожидания, идущие с верхних уровней вниз.

В нормальной ситуации результат восприятия будет зависеть как от активности рецепторов, на которые поступает информация, так и от активности нейронов верхних уровней, связанной с ожиданиями (предсказаниями на основе ранее полученной информации). При этом нейроны, отвечающие за взаимно согласующиеся гипотезы разных уровней, будут усиливать активность друг друга, т. е. входить в резонанс. Даже если какая-то гипотеза исходно и выглядела правдоподобнее, более слабые гипотезы, входящие в резонанс, в конечном итоге могут стать сильнее и выиграть. Именно поэтому нам кажется, что мы отчетливо слышим слова песни на фоне громкой музыки, когда хоть раз видели текст, даже если на память его воспроизвести не сможем. Если же текст нам не знаком (и к тому же он на иностранном языке), мы можем услышать что-то совсем другое и, поверив, что расслышали правильно, в последующие разы будем слышать то же самое.

Идея адаптивного резонанса позволила объяснить Гроссбергу некоторые особенности человеческого восприятия, например задержку в *осознании* сенсорной информации по сравнению со временем, требуемым для прохождения сигнала по зрительному или слуховому тракту. Эта задержка есть время, необходимое для установления резонанса и зависящее как от силы ожиданий, так и от степени неопределенности воспринимаемой информации.

С помощью адаптивного резонанса можно объяснить и тот факт, что на то, чтобы в первый раз увидеть объект, спрятанный на изображении-головоломке, уходит значительно больше времени, чем на его восприятие в последующие разы. На таких изображениях (из них наиболее известен рисунок далматинца) присутствует объект, складывающийся из каких-то других случайных объектов, но чтобы его увидеть, необходимы заметные усилия со стороны зрительной системы. Для распознавания спрятанного объекта требуется правильно сгруппировать видимые объекты, что требует большого перебора вариантов. Если бы обработка сенсорной информации шла строго снизу вверх, то зрительной системе пришлось бы каждый раз заново решать эту задачу. Второй раз взглянув на то же изображение, человек мог бы помнить, что на нем изображено, но не видеть этого до тех пор, пока нужная комбинация снова не нашлась. Однако распространение информации сверху вниз позволяет эффективно отсеивать неперспективные гипотезы нижних уровней, эффективно направляя поиск правильной интерпретации изображения. Здесь видна глубокая общность процессов восприятия и мышления. Влияние верхних уровней восприятия на процессы обработки информации на нижних уровнях еще больше видно по тому, что примитивные люди не подвержены некоторым оптико-геометрическим иллюзиям (т. е. иллюзиям, возникающим на уровнях восприятия до распознавания объектов), которым подвержены цивилизованные люди. Этот факт может быть, правда, объяснен не влиянием обратных связей, а изменением прямых связей под воздействием опыта.

В теории адаптивного резонанса проявляется проблема, которую часто называют дилеммой *стабильности-пластично-*

сти. В какой степени наши ожидания должны влиять на восприятие неоднозначной и недостоверной информации? И, наоборот, насколько только что полученная информация должна влиять на наши последующие ожидания? Последний вопрос напрямую связан с процессом обучения, который, конечно, находится в центре внимания в теории адаптивного резонанса. Ведь в этой теории строятся нейронные сети специальной архитектуры, для которых неизбежным остается вопрос об установлении связей на основе обучающей выборки. Непластичная сеть не сможет обучаться, поскольку не будет сохранять информацию, а сеть, целиком подстраивающаяся под текущие данные, тоже не сможет обучаться, поскольку не будет выполнять обобщения. Обе эти крайности интуитивно кажутся плохими, но как выбрать оптимум между ними? Нетривиальность этого вопроса видна по тому, что и люди при его разрешении руководствуются разными стратегиями (к примеру, консерватизм характеризуется максимальной стабильностью при минимальной пластичности). Эту дилемму затруднительно решить на примере конкретной архитектуры нейронной сети, поскольку, как мы увидим, она имеет фундаментальный характер и проявляется во всех методах машинного обучения.

Идея адаптивного резонанса многое объясняет. Однако при попытке ее реализации в форме ИНС возникает много подводных камней. К примеру, в этих ИНС информация хранится локально, поскольку каждой гипотезе должен соответствовать некоторый нейрон, что не соответствует известным принципам распределенного хранения информации в мозгу. Ведь не получается найти так называемые «бабушкины» нейроны, которые бы распознавали конкретные слова или образы. Этот термин возник около 1969 года благодаря высказыванию Джерома Литвина о том, что если такие нейроны существуют, то у него в голове должен быть и нейрон, который активируется при появлении его «бабушки». Хотя в некоторых экспериментах были найдены нейроны, реагирующие на определенные понятия, осталось неясным, активируются ли вместе с ними при этом какие-то другие нейроны и реагируют ли эти нейроны только на одно понятие или на какие-то другие тоже. Конечно, до сих пор нет доказательства и того, что «бабушкиных» нейронов нет. В конце концов, в первичной зрительной коре нейроны реагируют на конкретные стимулы, появляющиеся в определенном месте на сетчатке (так что представление изображений, по крайней мере на данном уровне, не является распределенным).

Проблема здесь, однако, не в самой распределенности, а в том, что общее число гипотез (например, число классов распознаваемых образов) ограничено числом нейронов. Для букв или даже отдельных слов это может быть допустимо — их все же не так много, и каждому из них можно поставить в соответствие свой нейрон. А вот число всех возможных предложений, состоящих лишь из пяти слов, уже превосходит число нейронов в мозге. То же относится и к интерпретации зрительных сцен, на которых может присутствовать бесчисленное множество комбинаций объектов.

Трудности с ИНС в теории адаптивного резонанса возникают не только для верхних уровней восприятия. Не решается и проблема инвариантного распознавания. Речь здесь идет лишь о разрешении неопределенности для зашумленных образов, тогда как даже простое смещение объекта препятствует распознаванию. Хотя, конечно, нельзя требовать от одной идеи решения сразу всех проблем.

Первоначальная простая архитектура ИНС в теории адаптивного резонанса впоследствии была сильно усложнена, и некоторые недостатки были устранены, однако принципиальные проблемы, связанные с отсутствием инвариантности и ограничением на число гипотез, решены не были. Хотя искусственные нейронные сети и позволили достаточно естественным образом воплотить идеи адаптивного резонанса, это удалось сделать лишь для весьма частного случая. Сама же идея адаптивного резонанса столь универсальна, что может применяться в решении любых NP-полных задач (а не только задач восприятия), которые разбиваются на слабо связанные подзадачи, где применение адаптивного резонанса позволяет снизить отрицательный эффект от принятия промежуточных решений. Если не ограничивать себя рамками ИНС, несложно представить себе такую реализацию адаптивного резонанса, в которой не будет столь жестких ограничений на множество гипотез. К примеру, на верхнем уровне может

«на лету» порождаться некоторое количество любых предложений, которые входят в резонанс с воспринимаемыми словами. Как отмечалось, в ИНС каждому предложению не может соответствовать собственный нейрон. И в то же время, если гипотезу представлять как совокупность нейронов, становится гораздо сложнее организовать резонанс между такими составными гипотезами на разных уровнях.

Можно сказать, что это одна из общих проблем для всех классических нейронных сетей: нейроны в них соединены попарно, поэтому сложно представить себе взаимодействие между группами нейронов, активность которых соответствует некоторому комбинаторному объекту, возникающему в конкретный момент времени, а не привязанному к некоторому нейрону. Это вызывает и трудности при обучении распознаванию подобных объектов, требующем синхронной модификации весов связей многих нейронов (поэтому обучение ИНС нередко производится централизованно внешним алгоритмом, а обучение на локальных правилах оказывается ограниченным). Конечно, мы рассмотрели далеко не все архитектуры ИНС. Но простого изменения архитектуры ИНС будет явно недостаточно для решения указанных проблем, характерных для любой модели, основанной на классических формальных нейронах.

НЕКЛАССИЧЕСКИЕ ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

К искусственным нейронным сетям с момента их появления отношение все время менялось: от полного неприятия до крайнего восторга. ИНС занимают значимое место в теории машинного обучения, однако их способность к обучению и сходство с биологическими нейронами сильно преувеличены. Исследователи, последовательно придерживающиеся бионического подхода (в рамках которого при создании технических систем заимствуются решения из живой природы), пытаются устранить оба недостатка, вводя в классическую модель формальных нейронов неучтенные в ней свойства

нейронов биологических. Поскольку последние несоизмеримо сложнее, вариантов модификации формальных нейронов может быть предложено очень много, и некоторые из этих модификаций уже получили заметное развитие.

Поразительно, насколько широко распространена в литературе информация о традиционных ИНС и насколько редко упоминается о существовании других нейросетевых моделей. Например, многие из тех, кто не занимается ИНС профессионально (и даже некоторые из числа профессионалов), не слышали о модели *спайковых сетей*.

Хотя достоверно известно, что распространение электрических сигналов между нейронами происходит в форме спайков (импульсов), частота которых зависит от уровня возбуждения нейронов, этот факт не учитывается при построении традиционных ИНС. В лучшем случае говорится, что в формальных нейронах с вещественным выходом (в отличие от нейронов Маккалока—Питтса с бинарным выходом) выходной уровень активности нейронов эквивалентен частоте спайков. При этом, однако, не моделируются другие динамические характеристики нейронов, например период рефрактерности. Кроме того, не ясно, играют все-таки какую-либо роль при кодировании информации индивидуальные интервалы между спайками, или важно только их общее количество. Существуют спайковые ИНС (и даже их аппаратные, а не только программные реализации), в которых моделируется динамика нейронных сетей на уровне отдельных импульсов (включая и период рефрактерности). При этом активность нейронов не пересчитывается синхронно (или в случайном порядке) для всех нейронов, как это обычно делается в классических ИНС, а изменяется непрерывно (не по тактам). Каждый импульс может привести к превышению порога возбуждения некоторого нейрона, поэтому точные времена прихода импульсов определяют моменты активации нейронов, в результате чего спайковые сети демонстрируют сложное динамическое поведение. Соединенные в пару нейроны могут возбуждаться попеременно, т. е. работать как осциллятор. В более крупных группах искусственных нейронов в спайковых сетях могут распространяться волны активности. Особый интерес представляет колебательная ак-

тивность в больших ансамблях нейронов в связи с широко известным фактом существования мозговых ритмов.

Почему информация — это обязательно то, что хранится в ячейках памяти, фиксируется в статичном состоянии системы? Циклическое изменение состояния нейронной сети может не только соответствовать постоянному изменению хранимой ею информации. Вместо этого информация может «храниться» в самой динамике изменения состояний, что делает информацию более «активной» (идея покажется не столь экстравагантной, если вспомнить о том, что информация часто передается с помощью разных волн). К примеру, кратковременная память человека, вероятно, реализована через динамическую активность нейронов (а не через веса связей между ними), и как только эта активность затухает, информация в памяти пропадает. Не менее интересны и хаотичные режимы работы таких ИНС, где информация «хранится» в странных аттракторах, к чему мы вернемся позже.

Сложное поведение интересно для моделирования, но как его использовать? Вопрос вовсе не простой и достаточно принципиальный: ведь и сами исследователи часто говорят о том, что спайковые сети на практике делают почти то же, что и классические ИНС, но требуют для этого гораздо больше вычислений. Спайковые сети не пользуются популярностью, поскольку их реализации на цифровых компьютерах не слишком эффективны. В этой связи удивительно упорство, с которым некоторые сторонники обычных ИНС говорят об отличии нейронных сетей от компьютеров, продолжая выполнять нейросетевое моделирование на этих самых компьютерах.

Наиболее естественное уточнение модели нейронов, учитывающее импульсный характер их активности, пока не привело к очередному прорыву в области машинного обучения. В этой связи стоит отметить большую разницу между точностью нейросетевых моделей и их полезностью для искусственного интеллекта. Ведь существуют гораздо более физически подробные модели единичных нейронов (еще в середине XX века Ходжкин и Хаксли вывели уравнения, описывающие распространение нервного импульса,

которые затем уточнялись) и целых нервных систем (например, упоминавшейся нематоды), но их, в основном, и не думают использовать при реализации технических систем. Эти модели используются преимущественно для воспроизведения откликов естественных нейронов. Один из самых амбициозных проектов по моделированию мозга — Blue Brain, реализующийся на суперкомпьютере Blue Gene, — включает попытку воспроизведения свойств нейронов чуть ли не до молекулярного уровня. Такая детальность, однако, вряд ли имеет смысл в технических приложениях. В этой связи нельзя сказать, что простейшие формальные нейроны хороши только тем, что в чем-то похожи на биологические нейроны. Видимо, важную роль играет то, в чем именно это сходство выражается.

Так какой же важный элемент нужно еще добавить в модель ИНС? Вспомним, какое именно ограничение в первую очередь проявлялось в ИНС. Это ограничение на возможности обучения, в особенности, обучения инвариантному распознаванию. При этом обучение ИНС в большинстве случаев весьма неестественно: например, очень сложно представить, чтобы в естественных нейронах обучающий сигнал распространялся обратно по аксону, как это представляется в алгоритме обратного распространения ошибки для обучения многослойного перцептрона, или тем более чтобы веса связей сети выбирались «на лету» в результате глобальной оптимизации. В довершение ко всему обучение выполняется (за редким исключением) не самими нейронными сетями, а внешними алгоритмами. Вдумайтесь в этот парадокс: искусственным нейронным сетям приписывают способность к обучению (в отличие от обычных алгоритмов), но обучают нейронные сети с помощью этих самых обычных (причем для каждой архитектуры ИНС разных) алгоритмов!

Подробные биофизические модели неплохо воспроизводят отклики ансамблей реальных нейронов при фиксированных связях между ними, но практически нет моделей, которые бы предсказывали, как будут меняться связи между нейронами в зависимости от подаваемых на них сигналов. Правило Хебба долгое время оставалось чуть ли не единственным правилом «внутреннего» обучения ИНС, имеющим при этом

хоть какое-то нейрофизиологическое правдоподобие. Если бы оно (или какое-то другое локальное правило) было общеприменимым и давало во всех случаях хороший результат обучения, то можно было бы назвать ИНС и самообучающимися, и принципиально отличающимися от обычных алгоритмов.

Идеи самоорганизующихся ИНС развивались в рамках концепции активных нейронов, в которой нейрон представляется не просто обработчиком информации, но самостоятельной системой со своими потребностями. Действительно, биологический нейрон — сложный, почти самостоятельный организм, который растет, развивается, оптимизируя свою деятельность так, чтобы получать больший объем питательных веществ. При этом в его ядре происходит экспрессия различных генов. Помимо электрических сигналов нейроны обмениваются многочисленными нейромедиаторами.

Если в моделях активных нейронов нейрон не проявляет активности, то он не будет получать питания, а значит, будет голодать и может даже умереть. Поскольку активность нейрона зависит от сигналов, приходящих от других нейронов, нейрон должен так настраивать свои связи, чтобы активироваться как можно чаще (или просто достаточно часто, чтобы быть «сытым»). При этом нейронная сеть управляет некоторым «телом», помещенным в среду, в которой оно может перемещаться, получать питание и т. д. Соответственно питание всей нейронной сети определяется тем, насколько адекватно поведение организма в целом, а для этого нейроны должны вырабатывать правильные управляющие воздействия как реакцию на определенные стимулы, поступающие на рецепторы. Нейроны для своего же блага должны обеспечить и инвариантное распознавание образов, и интеллектуальное поведение с решением проблем поиска, и т. д.

Идея очень красивая, и, на первый взгляд, может показаться, что она решает все проблемы. Но остается основной вопрос: как должны вести себя нейроны, чтобы правильно настроить свои связи? Попытка построения искусственной системы на основе данных принципов описана в книге 1990 года «Интеллектуальная квазибиологическая система:

Индуктивный автомат» Леонида Борисовича Емельянова-Ярославского, который, в частности, использовал опорный тезис: «разряд в нейроне нужен самому нейрону», считая, что «понять работу мозга можно, только ответив на вопрос: зачем мозг составляющим его нейронам?» — и пытался из слабоорганизованного множества нейронов, поставленных в условия ограниченного питания, получить нейронную сеть, обеспечивающую адекватное поведение организма в целом. На сходных принципах основана идея активных нейронов, описанная Александром Львовичем Шамисом в книге 2006 года «Пути моделирования мышления: Активные синергетические нейронные сети, мышление и творчество, формальные модели поведения и „распознавание с пониманием”» и дополненная некоторыми идеями синергетики. Причем, по заверению автора, именно эти идеи доведены до практической реализации в прикладных программах Графит, FineReader-рукопись и FormReader (ABBYY), хотя насколько в чистом виде они там применены, сказать трудно.

Как оказывается, из нейронной сети со случайными связями в результате самоорганизации нейронов не удается получить сети, обеспечивающей сложное поведение. Вернее, интеллект такой сети оказывается ограниченным интеллектом нейронов, от которых требуется быть слишком умными, например, уметь распознавать образы, различая паттерны сигналов, приходящих по разным аксонам, а не просто суммируя их. Конечно, естественные нейроны обладают очень сложным внутренним устройством. Но разве будет удивительным, если вся сеть научится распознавать образы, когда отдельные нейроны заранее умеют это делать?

Другой путь повышения возможности сетей из активных нейронов состоит во введении в них какой-то первоначальной структуры. Для прикладных задач эта структура будет зависеть от структуры входных данных, которые будут требовать очень специализированной предобработки. В случае же нейронной сети, управляющей поведением некоторого искусственного организма, эта структура будет включать специфические блоки памяти, эмоционального центра и т. д. Такое уточнение вполне естественно, ведь структура человеческого мозга вырабатывалась эволюционно на протяжении

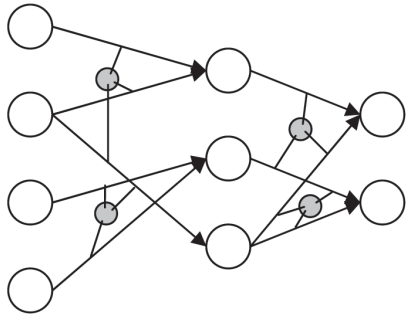
многих миллионов лет, поэтому странно было бы, если бы любая случайная нейронная сеть (или даже мозг другого животного) могла на протяжении своей «жизни» самоорганизоваться во что-то интеллектуальное. Но общая проблема от этого не устраняется: нейронная сеть должна заранее обладать способностью, например к распознаванию, благодаря либо сложности нейронов, либо начальной архитектуры; самообучение же оказывается весьма слабым.

Проблема здесь, возможно, в том, что как мозг не является однородной совокупностью нейронов, так и не бывает универсальных локальных правил обучения. Может, для каждой задачи, каждой архитектуры ИНС действительно нужен свой алгоритм обучения? Может, универсальный алгоритм обучения существует, но он слишком медлителен (как эволюция) и применим не для непосредственного обучения произвольных ИНС, а для создания частных алгоритмов обучения конкретных ИНС? В рамках теории ИНС ответить на эти вопросы не удастся. Но можно их сформулировать по-другому: могут ли в разных участках мозга работать разные алгоритмы обучения (формирования межнейронных связей)? Ведь существуют эксперименты, показывающие идентичность механизмов синаптической пластичности у самых разных животных — от моллюсков до человека. Но способности обучения человека гораздо шире способностей животных, поэтому человеческое обучение не может быть объяснено на уровне универсальных механизмов синаптической пластичности.

Во-первых, нейроны отличаются между собой морфологически и функционально (существует множество достаточно четко различающихся между собой типов нервных клеток), и даже тем, какие медиаторы они используют. Видимо, в таком делении должен быть глубокий смысл, но он остается пока не до конца понятым. Во-вторых, существуют глиальные клетки, о которых кратко упоминалось ранее. Напомним, что глиальных клеток на порядок больше, чем нейронов. До недавнего времени большинством ученых считалось, что глиальные клетки выполняют лишь вспомогательные функции, например, разделяют нейроны между собой и служат для них опорой. Было известно о том, что именно глия

обеспечивает питание нейронов. Но неужели основная часть (примерно 90 %) клеток в мозге не участвует в мышлении? В ряде недавних экспериментов было установлено, что между нейронами и глиальными клетками действительно идет двунаправленный обмен сигналами. Имеющиеся данные позволили создать модель трехстороннего синапса, образованного двумя нейронами (пре- и постсинаптическим) и *астроцитом* (зведообразной клеткой, относящейся к отдельному подтипу глиальных клеток). В отличие от нейронов астроциты передают медленно меняющиеся сигналы большой длительности (до десятков секунд) на малые расстояния. Эти сигналы могут влиять на проводимость трехстороннего синапса путем высвобождения дополнительных нейромедиаторов. Значит, вполне возможно, что глиальные клетки участвуют в обеспечении пластичности межнейронных связей, т. е. в процессе научения. Отсутствие описания именно этого процесса делает модели нейронных сетей столь неполными.

Предположения о роли глиальных клеток в обучении высказывались еще в 1960-е годы. Да и при построении моделей обучения ИНС некоторые ученые, в частности, автор когнитрона, ссылались на глиальные клетки. Однако в первом когнитроне потенциальная роль глии служила лишь для обоснования авторской модификации хеббовского правила обучения формальных нейронов. Сейчас же возникло отдельное направление исследований, посвященное искусственным нейроглиальным сетям (ИНГС), в которых глиальные клетки являются такой же частью сети, как и нейроны, что условно проиллюстрировано на приведенном ниже рисунке.



Условная схема нейроглиальной сети

Конечно, на этом рисунке отражены не все необходимые связи. Следовало бы добавить связи от нейронов к астроцитам и, возможно, связи между самими астроцитами. Такая модель позволит реализовывать наиболее сложные процедуры обучения. Пока еще, к сожалению, не ясны детали взаимодействия нейронов и глии, поэтому при построении моделей искусственных нейроглиальных сетей исследователи больше руководствуются необходимостью решения внутренних проблем ИНС, чем нейрофизиологической достоверностью.

Несложно представить себе, как поток информации через астроциты приводит к модификации весов связей между нейронами, что при правильно выбранной архитектуре ИНГС приведет к желаемому обучению. В частности, с помощью специально подобранной системы астроцитов многослойный перцептрон может быть расширен таким образом, чтобы он обучался методом обратного распространения ошибки без всякого внешнего алгоритма. Для ИНС с другой архитектурой можно подобрать свою систему астроцитов, чтобы обеспечить ее обучение.

Это очень важный для теории искусственных нейронных сетей результат. Ведь до этого ИНС не были самостоятельными, они опирались на традиционные алгоритмы именно в той части, в которой должны были их превосходить, — в обучении. Относительно ИНГС можно, по крайней мере, сказать, что они учатся сами.

На первый взгляд, все выглядит замечательно, поскольку в ИНГС может быть реализована любая процедура обучения. Но здесь возникает ряд сложных вопросов. На настоящий момент в ИНГС реализуются лишь известные алгоритмы обучения, причем перевод этих алгоритмов на язык системы астроцитов может быть нетривиальным, а разработка новых более эффективных способов обучения при этом не облегчается. Это отталкивает от ИНГС практиков, которые не видят, что же нового они дают, кроме как представляют по-другому уже известные вещи. Да и в теоретическом плане возникают определенные сомнения, поскольку получается, что в ИНГС каждой нейронной структуре должна соответствовать собственная специфичная глиальная структура, выполняющая

обучение. Конечно, если вдуматься, идея универсального простого правила обучения (типа правила Хебба) выглядит фантастичной, но все же она привлекательнее идеи того, что в каждом кусочке мозга работает свой специфический, заранее заданный, алгоритм обучения. Кроме того, в простейшей модели ИНГС сами алгоритмы обучения оказываются неизменными, поскольку нет элементов еще одного типа, которые бы настраивали связи астроцитов.

Можно, правда, представить себе нейронную сеть, в которой за обучение ответственны сами нейроны. Почему бы одним нейронам не подавать каких-то управляющих сигналов на другие нейроны? Реальные синаптические контакты столь сложны, что допускают и такое. Есть специальные нейроны, называемые *модулирующими нейронами*, которые образуют с другими нейронами контакты типа «синапс на синапсе». К примеру, хорошо известна роль этих нейронов в регуляции уровня болевой чувствительности. Давно известны (но почему-то остались без внимания среди разработчиков ИНС) такие эффекты, как, например, изменение рецептивных полей клеток зрительного тракта при резком *звуке*. Даже без знания того, какие механизмы это позволяют делать, сам факт быстрого временно́го изменения «весов связей» между нейронами говорит о слишком большой упрощенности традиционных ИНС.

Итак, прохождение сигнала через некоторые синапсы влияет не на активность нейронов, а на проводимость других синапсов! Это означает, что «веса связей» между нейронами могут меняться не только медленно в процессе обучения, но и динамически, под влиянием текущей информации. Но кто знает, насколько большую роль это играет в пластичности нервной системы? И сколько всего мы еще не знаем про нейроны! Ведь все, о чем шла речь, касалось лишь корректировки имеющихся связей. А как аксоны или дендриты нейронов растут и устанавливают «нужные» контакты? В ИНС же рассматриваются либо полносвязные сети, либо фиксированные архитектуры. Первое нереалистично для большого числа нейронов, а второе накладывает существенные ограничения на возможности обучения. Сейчас накапливаются данные о том, что глия также направляет и рост дендритов

нейронов, т. е. управляет не только проводимостью имеющихся синапсов, но и образованием новых связей.

Итак, классические ИНС имеют отдаленное сходство с биологическими нейронами, а для их обучения используются обычные алгоритмы. Простое подражание каким-то дополнительным свойствам биологических нейронов, неучтенным в классической модели ИНС, пока не позволяет достичь уровня самообучения ИНС, сопоставимого с естественными нейронами. Разные нейрофизиологи неоднократно проводили следующую аналогию. Представьте, что где-нибудь, скажем, на Марсе, оказался человеческий компьютер и марсиане пытались бы понять, как работает какая-нибудь программа, тыкая электродами в разные точки процессора. От физических принципов работы транзисторов до такого понимания простирается огромная пропасть. Даже выявив структуру связей в процессоре, марсиане не сильно приблизились бы к пониманию его предназначения (не то что выполняемой на нем программы, которую вряд ли имеет смысл описывать в терминах изменения электрического заряда в различных точках процессора!), особенно если на Марсе не принято использовать, скажем, булеву логику или арифметику. Лишь поняв смысл выполняемых операций (который, по крайней мере в случае компьютера, не вытекает напрямую из физических процессов, происходящих в транзисторах), можно хоть немного продвинуться дальше.

Если ИНС обладают какими-то полезными свойствами, нужно понять, почему они ими обладают. С годами утверждения о ИНС потеряли контекст и превратились то ли в рекламные лозунги, то ли в магические заклинания: «нейронные сети самообучаются, а компьютер действует по жесткой программе», «нейронные сети умеют распознавать, а компьютер — вычислять», «нейронные сети хранят информацию распределенно, а компьютер — локализованно» и т. д. С этими сопоставлениями можно было бы отчасти согласиться, если их отнести на счет существующего «программного обеспечения» мозга и компьютеров. Тогда можно было бы сказать: да, пока еще не существует программ, способных к неограниченному самообучению. Но часто делается подмена понятий: эти свойства приписываются ком-

пьютерам и нейронным сетям как таковым, т. е. различия объясняются на аппаратном уровне. Далее делается совсем нелепый (в контексте сопоставления компьютеров и нейронных сетей) переход: свойства естественных нейронных сетей переносятся на искусственные, работа которых моделируется на компьютере. Но ведь если ИНС исполняются на компьютере, значит, они реализуются каким-то алгоритмом; если при этом ИНС обладают способностью к обучению, значит, существуют самообучающиеся алгоритмы. Возможно, ИНС тут вовсе и ни при чем. К примеру, как отмечалось, распределенно представлять информацию в компьютере можно и без всяких ИНС — существуют же цифровые голограммы. С другой стороны, и в ИНС информацию можно хранить локально. Значит, различие между компьютерами и нейронными сетями более тонкое, и его нужно понять, чтобы эффективно использовать положительные качества ИНС и установить, как их развивать дальше.

НЕЙРОННЫЕ СЕТИ И МАШИНА ТЬЮРИНГА

Искусственные нейронные сети, состоящие из связанных простых «аналоговых» элементов, и машина Тьюринга, в которой движется головка, записывающая символы на ленту, выглядят совершенно непохожими. И тем удивительнее, что между ними есть глубокая общность. Это и составляет «неисчерпаемую» мощность понятия алгоритма — не конкретной его формализации в виде машины Тьюринга, нормальных алгорифмов Маркова или еще каком-либо из десятков существующих вариантов, а некоторого абстрактного понятия алгоритма, вмещающего в себя все эти частные формализации. Какое место здесь отведено искусственным нейронным сетям?

На первый взгляд кажется, что для любой искусственной нейронной сети несложно построить эквивалентную машину Тьюринга (или просто написать компьютерную программу, воплощающую эту сеть). Но верно ли обратное? В форме нейронной сети не так легко представить, скажем, алгоритм сортировки или алгоритм поиска подстроки в строке.

Вспомним, что с помощью нейронов МакКаллока—Питтса можно реализовать любую формулу логики высказываний или любой конечный автомат. Теорема Колмогорова об аппроксимации непрерывной функции от многих переменных суперпозицией конечного числа функций от одной переменной была использована для доказательства возможности реализации любой непрерывной функции (аргументы которой принимают значения от 0 до 1) с помощью перцептрона. К сожалению, конечные автоматы, как и все остальное, имеют гораздо меньшую мощность, чем алгоритмы.

Несмотря на это, часто говорят, что ИНС по своей алгоритмической мощности эквивалентны машине Тьюринга. Утверждение об эквивалентности требует существенного уточнения: если речь идет об ИНС *вместе с алгоритмом ее обучения*, то факт эквивалентности банален, поскольку может быть обеспечен алгоритмом, добавляемым к ИНС. А доказательство того, что для каждого алгоритма можно построить эквивалентную ему (необучающуюся) ИНС, достаточно сложно найти в литературе.

Если внимательно приглядеться, то можно увидеть, что для эквивалентности ИНС и машин Тьюринга для первых нужно предусмотреть потенциально бесконечный пул нейронов (поскольку машина Тьюринга отличается от конечного автомата, в первую очередь, бесконечной лентой). Тогда, если мы пытаемся сразу построить ИНС, эквивалентную некоторой машине Тьюринга (например, воплощающей сортировку массивов любых размеров), такую ИНС потребуется сделать фактически бесконечной — как по числу нейронов, так и по заранее заданным связям между ними. Это не очень красиво. В худшем случае нам бы хотелось сделать ее не бесконечной, а неограниченной. Тогда в нее бы подключалось столько нейронов, сколько нужно для конкретного набора входных данных: этот набор всегда конечен, но его размер заранее не задан, поэтому, если делать фиксированную ИНС любого конечного размера, найдется такой набор данных, который она обработать не сможет. Разница между фактически бесконечной ИНС и неограниченной ИНС очень существенна: последняя конечна для любого случая. Однако для нее нужен механизм подключения нейронов

(и установления с ними связей) в зависимости от объема входных данных. Если этого не сделать, то нейронные сети останутся конечными автоматами. Современные ИНС такой внутренней возможностью не обладают: их архитектура формируется внешним алгоритмом под конкретную задачу. Значит, проблема обучения ИНС тесно связана с проблемой их алгоритмической полноты!

Бесконечное число нейронов (для алгоритмической полноты) может быть заменено бесконечным числом состояния каждого нейрона. Неограниченная память формального нейрона может достигаться в предположении, что уровень текущей активности может представляться рациональным числом p/q, где p и q — натуральные числа. Неограниченное возрастание значений p и q при сохранении ограничения на значение их отношения создает возможность «хранения» в значении активности нейрона битовой строки неограниченной длины. Каждый нейрон тогда работает как машина с неограниченным стеком, что позволяет достичь алгоритмической полноты, но это выглядит слишком похоже на обычную машину Тьюринга.

Существуют свидетельства и того, что ИНС обладают неалгоритмическими возможностями (реализуют так называемые *сверхтьюринговы вычисления*). Этот математически строго доказанный результат может поразить воображение и заставить посвятить всю жизнь искусственным нейронным сетям. Однако при внимательном рассмотрении соответствующих доказательств обнаруживается, что «сверхтьюринговость» достигается, если величину активности нейрона считать вещественным числом, определяемым с *бесконечной точностью*. Если рациональному числу можно поставить бинарную строку конечной (но неограниченной) длины, то для точной записи произвольного вещественного числа потребуется бесконечная строка. При этом простое сложение двух вещественных чисел означало бы обработку бесконечно длинных бинарных строк, которая и приписывается нейрону.

Оба этих варианта (с рациональным и вещественным представлением уровня активности), хотя и представляют теоретический интерес, являются биологически и технически (и,

вероятно, физически) нереализуемыми, поскольку требуют неограниченной или бесконечной точности воспроизведения сигналов нейронов. Никакой сигнал, в том числе и аналоговый, не может нести бесконечного количества информации. Представьте, что уровень активности некоторого нейрона — это число с бесконечным количеством знаков после запятой. Любой минимальный шум будет приводить к тому, что в этом числе неизменным останется лишь конечное количество начальных знаков, а бесконечное количество последующих знаков случайным образом изменится. Поскольку нейроны для достижения сверхтьюринговости должны использовать всю бесконечную цепочку знаков, любой шум приведет к полному изменению функционирования сети, что сделает ее применение просто невозможным. Так что и в области искусственного интеллекта следует очень внимательно рассматривать корректные математические выводы на предмет их физической осмысленности. Конечно, можно вспомнить упоминавшуюся гипотезу физика Пенроуза о том, что в нейронах могут происходить неалгоритмические процессы за счет каких-то неизвестных квантовых эффектов. Но это пока не более чем гипотеза, которую не удается использовать на практике.

Есть еще один момент, отличающий ИНС от алгоритмов. Наиболее мощный результат Тьюринга заключался в том, что он показал существование универсальной машины — такой машины, которая может эмулировать любую другую машину по ее описанию. Именно это позволяет, в частности, реализовывать ИНС на компьютерах вместо того, чтобы для каждой ИНС или алгоритма создавать отдельное специализированное устройство. Несложно догадаться, что если для произвольного алгоритма сложно построить эквивалентную ему ИНС, то еще сложнее построить нейросетевой эквивалент универсальной машины Тьюринга, способный выполнять любой алгоритм. И, в частности, такая *универсальная ИНС* должна была бы уметь эмулировать действие любой другой ИНС по ее описанию! Даже просто представить себе ИНС, которая бы строила другие ИНС, гораздо сложнее, чем программу, которая бы строила другие программы.

Специалистам по ИНС проблема создания «универсальной ИНС» может показаться слишком надуманной и не имеющей «биологического аналога». Но здесь уместно вспомнить про теорию функциональных систем (ТФС) Петра Кузьмича Анохина. В данном контексте нас из ТФС интересует лишь то, что многие функции выполняются не фиксированными структурами мозга, а функциональными системами, динамически и избирательно организующимися из отдельных, не выбираемых заранее элементов. Каждая ИНС, напротив, обладает фиксированной архитектурой и выполняет определенную функцию.

Незамеченным следствием из ТФС Анохина является то, что мозг представляет своего рода «универсальную нейронную сеть», а значит, нужно думать, как изменить формализм ИНС, чтобы суметь с его помощью реализовать универсальную ИНС. По умолчанию полагается, что с использованием некоторых больших ансамблей нейронов это получится само собой. К сожалению, без конкретных механизмов управления связями между нейронами этого добиться будет не только сложно, но и вообще невозможно. Конечно, здесь мы не ответим на вопрос, что же это за механизмы должны быть. Но, по крайней мере, мы уточнили проблему «внутреннего» обучения ИНС: теория типа искусственных нейроглиальных сетей должна дать универсальную ИНС, для чего, возможно, придется не только ввести специальные глиальные клетки, но и уточнить модель формального нейрона.

Как уже говорилось, самим нейронным сетям приписывают способность распознавать в отличие от компьютера. Но ведь не любая ИНС выполняет распознавание, да и не любую задачу распознавания легко решить с их помощью. Представьте, вам дают пары чисел в качестве обучающей выборки: (3751, 1357), (4382, 2348), (7289, 2789), которые образуют некоторый класс. А потом просят «распознать», какие пары из перечисленных принадлежат этому же классу: (2384, 2023), (3891, 1389), (5261, 1093), … .

Мозг человека сформирует «алгоритм» связи между числами в правильных парах, причем сделает это очень быстро. А теперь попробуйте построить ИНС, которая сумела бы сделать хоть что-то похожее. Кто-то может назвать этот пример неудачным и сказать, что «распознавание» здесь выполняется на более высоком уровне, чем распознавание,

скажем, лиц. Как будто эта отговорка решает проблему, и достаточно сказать, что сложные операции выполняются сложными нейронными сетями! И вряд ли можно утверждать, что распознавание лиц выполняется простой нейронной сетью. Очевидно, распознают не нейронные сети как таковые, а конкретные их архитектуры, воплощающие определенные процедуры. И нельзя думать, будто бы, взяв ИНС посложнее, можно заставить ее учиться чему угодно. Если это не будет «универсальная ИНС», то можно заранее предсказать, что она не сможет учиться произвольным процедурам, выявлять непредусмотренные заранее закономерности.

Как это ни странно прозвучит, мозг будет слишком негибким по сравнению с компьютерами, если он не будет обладать свойством, аналогичным УМТ, — уметь «виртуально исполнять» нейронную сеть (или просто алгоритм) по ее «описанию». Ведь в этом случае он будет уметь делать лишь то, что заложено в связях между его нейронами, изменением которых он сам как целое не управляет. Возможно, без этого нельзя будет достичь и эффекта самосознания. Итак, свойство УМТ фундаментально, и оно пока не реализуется для ИНС, которые оказываются лишь частным случаем алгоритмов. Создать универсальную ИНС, которая бы могла эмулировать любую конкретную ИНС по ее описанию, очень непросто, и с использованием классических формальных нейронов этого пока никому не удавалось. Конечно, эту задачу можно сильно облегчить, если усложнить нейроны и передаваемые ими сигналы. К примеру, можно ввести идентификаторы нейронов так, чтобы любой нейрон мог передавать информацию любому (не связанному с ним) нейрону по его идентификатору. И такие модификации ИНС действительно есть, однако они слишком приближены к обычным алгоритмам. Но стоит ли тогда их интенсивно исследовать, если это свойство уже давно реализовано в компьютерах?..

Ведь идея компьютеров тоже развилась из попытки описания мышления человека, т. е. у них уже есть некое функциональное подобие мыслительным процессам. ИНС воплощают лишь одну из особенностей биологических нейронов, характерную не только для мозга человека, но и для какого-

нибудь червяка. Почему именно эта особенность считается важной? Почему игнорируется море других особенностей?.. В этом смысле видно, что концепция алгоритмов является гораздо более фундаментальной, чем концепция ИНС. В конце концов, ИНС — это просто одна из форм представления алгоритмов. И мыслить об ИНС следует как о частном (но, возможно, важном) способе представления алгоритмов. Соответственно ИНС нужно противопоставлять не алгоритмам вообще, а другим их частным представлениям.

Все же ИНС как способ представления алгоритмов (алгоритмов в том самом обобщенном смысле, а не в качестве их какой-то конкретной архитектуры) обладает рядом уникальных особенностей. Бесспорное, хотя и не вполне уникальное преимущество ИНС заключается в удобстве перенесения на цифровые вычислители с массовой параллельностью, что позволяет создавать высокопроизводительные специализированные нейрокомпьютеры. В то же время распараллеливание алгоритмов, представленных в традиционной форме, — сама по себе нетривиальная задача. Но это, скорее, технический момент, поскольку в алгоритмическом плане (или в плане решения новых задач) такая параллельность дает немного. Для нейрокомпьютеров закономерным образом проявляется и то, что ИНС многих типов обучаются «внешним» алгоритмом. Как алгоритм обучения заложить в нейрокомпьютер? Нужно ли к нему подсоединять для этого обычный процессор? Интересной была бы разработка нейропроцессора для нейроглиальных сетей.

Последовательные сторонники «неалгоритмического» подхода рассматривают не цифровые, а аналоговые способы реализации ИНС. Например, еще полвека назад были популярны оптические вычисления, интерес к которым сохраняется и сейчас. Работу таких физических ИНС нельзя *точно* описать в форме алгоритма. Возможно, за этим кроются какие-то возможности, но пока эта неточность является недостатком не алгоритмов, а физических моделей ИНС. Можно провести такую параллель: передачу телевизионного сигнала аналоговым способом нельзя точно описать алгоритмически; но неточности описания будут связаны с шумами в аналоговом сигнале; невозможность точного предсказа-

ния этих шумов является сомнительным недостатком, ведь при передаче телевизионного сигнала цифровым способом ставится задача не имитации сигнала, как если бы он передавался аналоговым способом, а задача воспроизведения исходного сигнала без искажений. И именно последняя задача цифровой передачей решается гораздо лучше, чем аналоговой. В этом смысле сомнительно, что аналоговые процессы в мозгу следует воспроизводить аналоговыми системами. Возможно, цифровые системы с теми же задачами смогут справиться лучше, чем сам исходный аналоговый прототип — мозг. При этом и мозг часто старается использовать дискретные представления информации.

Так в чем же особенность ИНС как способа представления алгоритмов? Хоть и простое, но важное соображение заключается в том, что функционирование нейронной сети обычно мало меняется при небольших изменениях в ее строении. Конечно, это свойство ИНС хорошо известно. Однако оно обычно отождествляется со способностью мозга продолжать корректно работать при не слишком значительных повреждениях. При этом говорится, что изменение даже одного байта в компьютерной программе часто приводит к необратимым нарушениям в ее работе. Конечно, это так, но высокая устойчивость к повреждениям вряд ли напрямую связана с интеллектуальностью, а на практике для программ она бывает нужна нечасто и может достигаться иными способами.

Однако такая устойчивость крайне важна по совсем иным соображениям. Дело в том, что обучение можно представить как поиск оптимальных алгоритмов решения каких-то задач. В обычном представлении алгоритмов как программного кода такой поиск осуществлять трудно. Представьте, что вы написали какую-то программу на традиционном языке программирования. А потом из нее случайным образом было удалено или заменено несколько символов. Что будет с этой программой? Скорее всего, она просто перестанет компилироваться или, если повезет, будет просто работать совсем по-другому. Малые изменения в программе приводят к большим изменениям в ее работе. Зная, как нужно изменить работу программы, сложно «угадать», как для этого нужно изменить саму программу.

Программист работает не только с кодом, но и с каким-то глубинным представлением алгоритмов. ИНС же в большем числе случаев позволяют приблизиться к искомому алгоритму маленькими шажками путем постепенного изменения весов связей. Более формально можно сказать, что ИНС задают более гладкую метрику в пространстве алгоритмов, что и облегчает проблему их оптимизации (обучения). В этом смысле дополнительный интерес представляют нейроглиальные сети, поскольку в представлении ИНГС облегчается проблема автоматического поиска самих алгоритмов обучения, задающихся сетью астроцитов.

В традиционные программы тоже могут вводиться компоненты (например, параметры процедур), допускающие плавное изменение, но это делается человеком для каждого конкретного случая, тогда как в ИНС эта «способность» заложена сразу, чем и вызвана большая эвристическая сила ИНС в области машинного обучения. К сожалению, при этом само представление многих алгоритмов в форме ИНС оказывается значительно более громоздким, в связи с чем мало кто «программирует» на языке ИНС. Возможно, существуют представления алгоритмов, одновременно эффективные и гибкие. Но с такой точки зрения данную проблему почти никто еще не рассматривал.

Искусственные нейронные сети не отвечают на сложные вопросы машинного обучения, причины чего становятся ясны при рассмотрении ИНС как частного представления алгоритмов. Тогда традиционные алгоритмы должны предоставлять не менее широкие возможности в этой области, в связи с чем они также заслуживают изучения.

РАСПОЗНАВАНИЕ И ИНДУКЦИЯ

РАСПОЗНАВАНИЕ ОБРАЗОВ И ОБУЧЕНИЕ

Восприятие и обучение — две тесно связанные задачи. Наиболее явно они соединяются на этапе распознавания. Не слишком большой ошибкой будет предположение о том, что

чивается воплощенностью интеллекта (его связью с внешним миром через сенсорику и моторику). Однако рассмотрение высших когнитивных операций само по себе вряд ли даст решение проблемы сильного ИИ. Более последовательный подход заключается в том, чтобы начинать с прообраза мышления как поиска в физическом пространстве (не требующего поддержания модели мира) и расширять его до мыслительных процессов символьного уровня постепенно. В этой связи основной интерес могут представлять исследования моделей адаптивного поведения животных и его возникновения в ходе эволюции.

Часть четвертая

СТАНОВЛЕНИЕ ИНТЕЛЛЕКТА

ИНТЕЛЛЕКТ И ЭВОЛЮЦИЯ

АНИМАТЫ

В когнитивной робототехнике основное внимание уделяется высшим познавательным функциям. Хотя по сравнению с классическим ИИ, в котором моделирование интеллекта начиналось с уровня сознания, здесь затрагиваются и более низкоуровневые механизмы, они выступают в качестве ненадежных промежуточных блоков. Нельзя ли все же начинать строительство здания ИИ не с крыши (сознания), а с фундамента — тех неосознаваемых процессов, которые постепенно формировались в ходе эволюции для обеспечения выживания животных и по отношению к которым сознание является лишь наиболее поздней надстройкой?

В таком ракурсе проблема ИИ рассматривается в рамках области исследования, называемой *«Аниматы»* (или, по-другому, *«Адаптивное поведение»* — АП). Название «анимат» происходит от соединения слов «animal»+«robot», т. е. «роботы-животные», под которыми подразумеваются некоторые искусственные (модельные) организмы, живущие в реальном или виртуальном мире. Иногда, абстрагируясь от связи с животными, исследователи называют свои детища просто *интеллектуальными агентами*.

Большое внимание здесь уделяется конструкции актуаторов (исполнительных устройств). В первую очередь решается проблема перемещения анимата, без которого затруднительно выполнять другие важные функции — взаимодействия

с предметами, обеспечения выживания (получения энергии), а также социального взаимодействия. Наиболее простым решением (с точки зрения координации движения) является использование колес или гусениц, но это решение не позволяет эффективно перемещаться по сложной местности, например по лестницам. В связи с этим популярным является воспроизведение способов движения живых организмов. Интересный способ заимствуется у насекомых: использование шести ног при перемещении позволяет сохранять опору на три из них, что обеспечивает постоянное статическое равновесие. Более высокоразвитые животные обходятся меньшим числом ног, поэтому решения с четырьмя и двумя ногами представляют интерес в рамках АП. В случае четырех ног сохранение постоянного статического равновесия подразумевает отрывание от земли лишь одной ноги за раз, что приведет к очень медленному перемещению. Еще сложнее (и тем интереснее) движение на двух ногах: в этом случае, даже стоя на месте, приходится поддерживать равновесие, реагируя всем телом, ведь любое небольшое внешнее воздействие может привести к падению. Не менее интересными представляются попытки моделирования других способов перемещения, т. е. попытки научить роботов ползать, плавать, летать. Жестко запрограммированные движения будут чувствительны к любым неровностям. Строго описать все ответы на все возможные воздействия вряд ли возможно, особенно в условиях неполноты информации, поэтому вопросы адаптивного управления при организации движения оказываются крайне важными. Однако адаптация здесь обычно выполняется локально и на основе специфических эвристик. Еще более интересным является само адаптивное поведение: что оно собой представляет, какова должна быть архитектура систем управления, чтобы они могли обладать способностью приспосабливаться к изменяющейся внешней среде, и т. д.

Традиционные роботы, разрабатываемые в конкретных прикладных целях (и даже снабжаемые при этом когнитивными архитектурами), не демонстрируют такого же уровня адаптивности, как и животные. Недостаточная адаптивность проявляется в неспособности к автономному функционированию в заранее неизвестных условиях. С этим легко смириться, если робот удовлетворительно решает поставленную перед ним задачу. К примеру, то, что робот-пылесос может где-то застрять или заблудиться без возможности вернуться на базу для подзарядки, является неприятным, но не смертельным: владелец его спасет и даже облегчит ему работу, убрав разбросанные на полу вещи. То, что для большинства роботов лишь несущественно снижает потребительские качества, для животных означает неминуемую гибель. Многие ученые предполагают, что адаптивное поведение тесно связано с интеллектом, в связи с чем для ИИ моделирование пусть простых, но полностью самостоятельных организмов (аниматов) гораздо важнее, чем моделирование высших когнитивных функций или создание полезных, но узкоспециализированных роботов.

Официально направление «Адаптивное поведение» сформировалось в 1990 году после проведения в Париже Первой международной конференции «Симуляция адаптивного поведения (от животных к аниматам)». Стоит отметить, что это направление фактически существовало и раньше, в чем несложно убедиться по материалам книги М. Г. Газе-Рапопорта и Д. А. Поспелова «От амебы до робота: модели поведения», опубликованной в 1987 г. В качестве одной из ранних работ в этой области можно привести проект «Животное» под руководством советского кибернетика Михаила Моисеевича Бонгарда, проводившийся еще в 1960–1970-х годах, но по структуре соответствующий современным проектам в области АП.

Сегодня подобные исследования, однако, настолько более многочисленные, что даже сложно выбрать несколько основных или наиболее показательных проектов. Моделируются самые разнообразные животные: от насекомых (и даже червей) до обезьян. Цель создания аниматов тоже может быть самой разной: от разработки коммерческих домашних роботов-животных до изучения принципов адаптации в самом общем виде. Наиболее последовательными в рамках данного направления являются исследования простейших организмов, снабженных минимумом априорной информации об окружающем мире.

Представим себе «одноклеточного» анимата, перемещающегося на плоской поверхности в поисках пищи. Каким должен быть оптимальный алгоритм перемещения? Конечно, можно сказать, что, не зная свойств мира (т. е. не имея его модели), нельзя вывести оптимальный алгоритм. Но в том-то и проблема, что требуется организовать поведение анимата в заранее неизвестных условиях. Можно предложить некоторый регулярный способ перемещения, например, по расширяющейся спирали. Казалось бы, это максимально повысит вероятность нахождения пищи. Но легко представить себе мир, в котором пища распределена неравномерно, «островками». Тогда более эффективным будет алгоритм перемещения по прямой до первого попавшегося «островка» с последующим его обходом: анимат движется прямо, когда голоден, и по спирали, когда сыт. Однако можно представить себе и мир, в котором «пища» убегает от прямолинейно перемещающегося анимата, и ее можно достичь, только совершая резкие повороты. Для любого фиксированного алгоритма перемещения можно придумать мир, в котором этот алгоритм будет неэффективен. Более того, из всех возможных миров таковых будет большинство. Истинная неопределенность и заключается в том, что свойства внешнего мира могут оказаться любыми. Как не допустить того, чтобы алгоритм перемещения при этом оказался полностью непригодным?

Вопрос этот не столь праздный, ведь современные коммерческие роботы, не предназначенные для работы в детерминированных условиях, регулярно сталкиваются с непредвиденными разработчиками ситуациями. Чем жестче программа перемещения, заложенная в робота, тем вероятнее, что непредвиденная ситуация станет для нее фатальной.

Неопределенности окружающего мира можно лишь противопоставить разнообразие собственного поведения. Тогда простейший анимат должен просто совершать случайные действия. Когда никакой априорной информации нет, ничего другого и не остается. Этот рецепт кажется слишком простым. Чем он отличается от фиксированной программы? Посмотрим на цепочку действий, порожденных фиксированной программой. Ее алгоритмическая сложность будет

ограниченной. В то же время алгоритмическая сложность цепочки, состоящей из случайных действий, будет постоянно увеличиваться с ростом длины цепочки (здесь принципиально наличие «неалгоритмического» источника хаоса). В такой цепочке можно найти любую подцепочку, поэтому при случайном выборе действий существует вероятность (хоть и, возможно, очень низкая) выжить в любом мире, в котором выживание в принципе возможно. Если вернуться к примеру с перемещением простейшего анимата на плоскости, то несложно увидеть, что случайные перемещения будут приемлемы для любого из рассмотренных вариантов размещения пищи.

В самом начале, при обсуждении лабиринтной гипотезы мышления, мы отметили важность перебора вариантов как одной из основ мышления. Однако эвристическое программирование развивалось преимущественно на примерах детерминированных формальных миров. Хотя мы отмечали беспорядочность поведения животных, встречающихся с неожиданным затруднением, из этого был сделан лишь вывод о том, что животными выполняется перебор вариантов. Теперь мы видим, что случайность этого перебора имеет принципиальное значение в условиях неопределенности.

Конечно, совершение просто случайных действий оказывается не самой эффективной стратегией и в простейших случаях. Совмещать регулярное и случайное поведение можно в рамках разных архитектур. В том числе могут использоваться и ИНС. Оказывается, даже простейшие, состоящие всего из нескольких нейронов (с источником случайной активности), сети способны обеспечивать интересное поведение. Пусть, к примеру, каждому из возможных действий поставлен в соответствие свой нейрон. Эти моторные нейроны соединены отрицательными связями и подавляют активность друг друга так, что лишь один из нейронов остается активным. Из-за случайного возбуждения активность нейронов со временем меняется. Если случайное возбуждение очень велико, то последовательность действий будет полностью случайной. В противном случае будут наблюдаться цепочки повторяющихся действий некоторой длины (например, движение в одном направлении на некоторое расстояние

с последующей случайной сменой направления). Если в качестве действия выступает поворот, то, напротив, длительная активность одного нейрона будет означать более высокую кривизну траектории. Далее может быть добавлен сенсорный нейрон, который регулирует частоту переключений активности моторных нейронов в зависимости от присутствия пищи. Тогда в целом перемещения будут случайными, что обеспечит возможность анимату выходить из тупиковых ситуаций, но участки без пищи будут проходиться по более прямолинейным траекториям, что повысит эффективность поиска (если такая регулярность поведения не будет противоречить свойствам мира).

Подобная модель была, в частности, описана в работе В. А. Непомнящих «Модели автономного поискового поведения» (сборник «От моделей поведения к искусственному интеллекту»). Автор также отмечает, что сходное поведение проявляют многие животные, в частности, личинки златоглазки двигаются по слабо искривленным траекториям, пока не найдут тлю, которой питаются. После обнаружения одной особи тли кривизна траектории златоглазки возрастает для более детального обследования окрестностей, где ожидается присутствие других особей колонии тли. Аналогичное поведение проявляют и личинки ручейника, старающиеся на дне водоема найти наиболее подходящие частички для строительства «домика» вокруг себя.

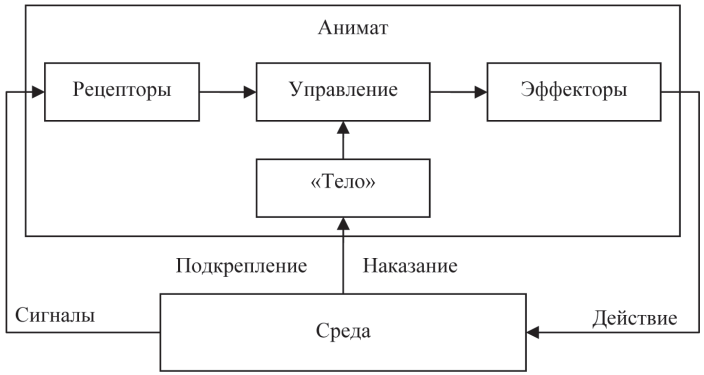
Конечно, простейшие нейросетевые модели отнюдь не решают проблему адаптивного поведения. И дело здесь не в том, что из-за малого числа нейронов модель не показывает всего разнообразия форм поведения, доступных даже личинкам насекомых. Проблема заключается в том, что адаптивность здесь строится на основе априорной информации о мире (хотя и включает локальный учет информации, поступающей от сенсора). Хотя нередко и утверждается, что в подобных системах управления не используются никакие модели внешнего мира или даже сведения о нем, но неявно они, конечно же, в архитектуре ИНС содержатся. К примеру, что лучше для ручейника при строительстве домика: собирать частички найденного размера и формы или продолжить искать более подходящие частички? Для

ответа на этот вопрос нужно знать вероятности нахождения тех или иных частичек. Кто-то скажет, что ручейник их «не знает». Но ручейник реализует поведение, которое эквивалентно использованию вполне конкретных вероятностей. Конечно, в отдельных поведенческих актах может проявляться неопределенность в значениях этих вероятностей, так что в идентичных ситуациях ручейник будет вести себя по-разному, но все же в среднем выбор ручейника будет отвечать некоторым вероятностям. Если анимата, управляемого описанной ИНС, поместить в среду с другими свойствами, то его поведение будет менее эффективным, чем поведение с полностью случайным выбором действий. Настоящая адаптивность подразумевает, что связи в такой сети должны настраиваться на основе опыта.

Естественно, без решения фундаментальных проблем машинного обучения адаптивность поведения аниматов, будь оно основано на нейросетевых или каких-либо других технологиях, останется существенно ограниченной. Но даже без решения этих фундаментальных проблем исследование адаптивного поведения позволяет выявить некоторые дополнительные важные аспекты интеллекта.

В наиболее простом виде взаимодействие анимата со средой описывается представленной ниже схемой. Анимат обладает набором эффекторов, с помощью которых может совершать действия, а также набором сенсоров, получающих информацию о мире. Далее ставится вопрос о конкретизации этих общих блоков (особенно системы управления). Такое уточнение может браться как из работ по исследованию естественных систем адаптивного поведения (к примеру, в этих целях широко используется теория функциональных систем Анохина), так и подбираться искусственно для проверки какой-либо идеи. В качестве примера такой идеи можно привести проверку значимости мотивационных центров, блоков планирования или долговременной памяти.

Приведенная ниже схема не отражает временных характеристик поведения, но понятно, что динамика взаимодействия с внешним миром, являющимся неограниченным источником неопределенности, представляет наибольший интерес. Тем не менее уже на этой схеме видно, что поми-



Простейшая схема анимата

мо рецепторов и эффекторов аниматы должны обладать еще одним каналом для получения «информации» из внешнего мира — «телом», через которое анимат узнает об успешности своего поведения.

В простейшем случае от «тела» используется только один канал «хорошо/плохо» (по нему могут передаваться как бинарные, так и вещественные значения). Полезность этого канала сложно переоценить. Благодаря нему становится возможным *обучение с подкреплением*. В рамках этого обучения строятся такая модель мира и стратегия поведения, которые позволяют максимизировать целевую функцию, задаваемую средой. К обучению этого типа можно, в частности, отнести формирование условного рефлекса. Проблема здесь, однако, в том, что подкрепление или наказание могут заметно отстоять во времени от вызвавшего их действия. В связи с этим в теории обучения с подкреплением приходится рассматривать более сложные по сравнению с условным рефлексом методы. В следующем по сложности случае (после случая безусловного стимула, непосредственно следующего за выполненным действием) количество состояний внешнего мира и количество возможных действий считается конечным (и небольшим). Тогда удается оценить вероятности перехода мира из одного состояния в другое в зависимости от последовательности совершаемых действий, а также узнать, какие состояния мира соответствуют подкреплению и наказанию. На основе этих данных уже можно выработать правила выбора действия в зависимости от состояния мира.

Конечно, даже в простых игровых мирах число состояний невообразимо велико. Что уж говорить о реальном мире? Для применения методов данного типа нужно как-то обобщать как состояния мира, так и собственные действия (проблема такого обобщения уже кратко обсуждалась нами на примере формирования условных рефлексов). Как в природе животным удается успешно обучаться на основе подкрепления? Конечно, обучение сложным навыкам идет в стиле обучения с учителем. Для этого в частности широко используются специальные методы подражания, которые совместно с другими видами социального взаимодействия сейчас стало популярным моделировать на базе аниматов и когнитивных роботов. Однако исходно эти навыки должны вырабатываться при обучении с подкреплением. Помимо большей эффективности самих методов обучения, детально проработанных в ходе эволюции, важную роль могут играть такие «телесные» механизмы, как потребности, мотивация, эмоции.

ИСКУССТВЕННЫЕ ЭМОЦИИ

Откуда у животных взялась способность испытывать боль и удовольствие? С одной стороны, эта способность выглядит крайне естественной и полезной для выживания. Однако при проектировании аниматов становится понятно, что сама собой она не появляется. Можно провести следующую аналогию: при игре в шахматы известны правила поражения, но из них явно не следуют способы оценки качества текущей позиции. Чтобы понять, что потеря фигуры в данной игре обычно является нежелательной, нужно сыграть много игр (чтобы проследить, что именно потери фигур вели к проигрышу) или использовать какие-то априорные знания. Также и в случае с животными: без использования достаточно глубоких знаний о природе сложно сказать, какие ситуации будут приближать или отдалять смерть живого организма. Боль и удовольствие следует рассматривать как общие эвристики оптимизации поведения, выработанные в ходе эволюции.

Однако в силу своего обобщенного и локального характера они далеки от универсальности, что можно сравнить с такой базовой эвристикой, как суммарная сила фигур в некоторой игре. Потеря фигуры — боль, съедание — удовольствие. Такая простая оценивающая функция несравненно лучше, чем полное отсутствие каких-либо эвристик, но ее все же недостаточно.

Вполне естественно, что в ходе эволюции были выработаны и более сложные эвристики. К таковым можно отнести эмоции и чувства, которые не столь локальны, как боль и удовольствие. Биологическая роль ощущений, эмоций и чувств вполне понятна: голод заставляет заранее искать пищу, страх позволяет избегать опасности, удивление — направлять внимание на новую информацию и так далее. Как в шахматах оценивающая функция состоит из слагаемых, включающих не только силу фигур, но и позиционное преимущество, и эта функция подбирается так, чтобы ее максимизация с наибольшей вероятностью приводила к выигрышу, также «взвешенную сумму» эмоций и чувств можно трактовать как оценивающую функцию, в которую эволюционно заложены эвристики выживания. С такой ролью эмоций (хотя и в менее «кибернетической» формулировке) вполне согласны современные психологи, в частности, в книге «Эмоции и чувства» профессора Е. П. Ильина можно прочитать:

> «Оценочная роль эмоционального реагирования вместе с развитием нервной системы и психики живых существ видоизменялась и совершенствовалась. Если на первых этапах она ограничивалась сообщением организму о приятном или неприятном, то следующей ступенью развития явилась, очевидно, сигнализация о полезном и вредном, а затем — о неопасном и опасном и, наконец, более широко — о значимом и незначимом. Если первая и отчасти вторая ступень могли обеспечиваться только таким механизмом эмоционального реагирования, как эмоциональный тон ощущений, то третья ступень требовала другого механизма — эмоции, а четвертая — чувств (эмоциональных установок)».

Однако нужно ли все это искусственному интеллекту? Должен ли и может ли ИИ испытывать чувства, эмоции, быть способным любоваться природой или воспринимать

прекрасное, понимать чувства людей, юмор, поэзию? Стереотипным является представление о гипотетическом искусственном разуме как о бездушной машине, действующей в соответствии с холодной логикой. Этому стереотипу предшествует традиционное противопоставление разума и чувств. Действительно, зачем в искусственном интеллекте пытаться воспроизводить то, что входит в противоречие даже с естественным интеллектом? Ведь в психологии долгое время за эмоциями признавалась лишь дезорганизующая функция, ослабляющая здравый смысл. Кроме того, эмоциональные интеллектуальные системы сейчас редко нужны на практике (сложно представить себе робота на производстве, плачущего из-за превышения нормативного показателя бракованных изделий). Лишь персонажей в компьютерных играх имеет смысл наделять «эмоциями» для «оживления» общения между ними и человеком. Не слишком-то широкая область приложения для столь туманной проблемы, как искусственные эмоции! Тем не менее эта проблема будет становиться все более актуальной при массовом распространении роботов. Уже сейчас в коммерческие роботы-животные и коммуникационные роботы закладываются механизмы демонстрации эмоций домашних питомцев или человека (с помощью движений и мимики). Должны ли эти роботы лишь имитировать проявления эмоций или в самом деле их переживать?

Естественно, поверхностная имитация эмоций не даст роботам их понимания. Как уже было показано, понимание существительных и глаголов требует их семантического обоснования в сенсорах и эффекторах. Также и понимание эмоций и чувств требует соответствующей семантической основы в собственных переживаниях. Как эмоции можно встроить в структуру ИИ?

В простейшем случае эмоции моделируют конечным автоматом, состояния которого соответствуют разным эмоциям. Этот автомат содержит набор правил перехода из одного состояния в другое, а также прямые и обратные связи с основным интеллектом (системой управления) и с интерорецепторами (боль, удовольствие). Упрощенная схема автомата представлена ниже. В более сложном случае из каждого состояния должны исходить связи, маркированные различ-

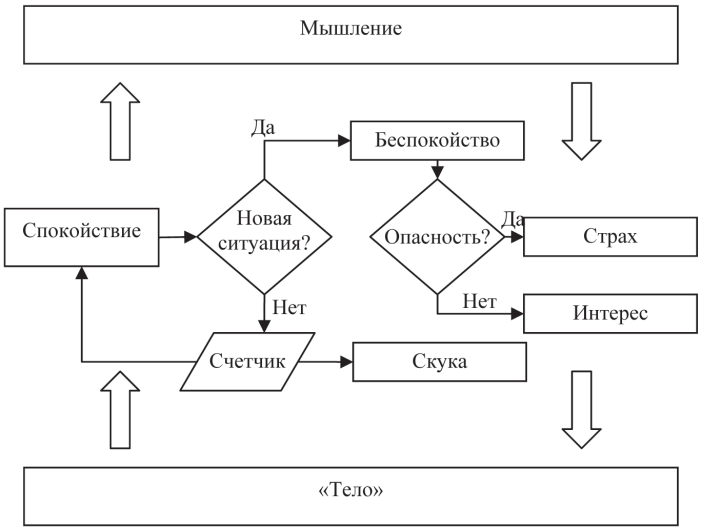


Схема автомата, переключающегося между несколькими эмоциональными состояниями

ными событиями. В тех же целях можно использовать не конечный автомат, а нейронную сеть.

На подобных моделях несложно понять, зачем человеку и животным так много разных эмоций (что вызывает удивление у некоторых психологов). Если представить, что животным осуществляется оптимизация оценивающей функции, представленной как сумма некоторого количества признаков, то каждая эмоция будет отвечать своему признаку. Увлекательной и не слишком сложной является задача создания системы, переводящей компьютерный интеллект в разные эмоциональные состояния так, чтобы это соответствующим образом сказывалось на его поведении. Несложно (на примитивном, конечно, уровне) также промоделировать такие личностные характеристики, как храбрость, любознательность и т. д. Для этого достаточно просто в оценивающей функции выбирать разные коэффициенты перед соответствующими эмоциями.

Переход эмоциональной системы в новое состояние вполне может переводить в другой режим работы и мышление. Тогда эмоции служат эвристиками, или обобщенными признаками ситуаций, позволяющими распознавать, какая стратегия поведения в текущий момент наиболее приемлема.

Подобную точку зрения высказывал Марвин Минский в книге «The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind» («Эмоциональная машина: Логика здравого смысла, искусственный интеллект и будущее человеческого разума»), одной из немногих книг в области ИИ, посвященных этой проблеме. Минский утверждал, что каждое эмоциональное состояние — это просто отдельный стиль мышления, и эмоции, в отличие от распространенного мнения, не отличаются принципиально от рационального мышления. Ведь, как отмечалось выше, даже ученые дают своим теориям эмоциональную оценку. Конечно, механизмы переключения между эмоциями должны обладать собственной «логикой». И то, что лимбическая система, в том числе участвующая в формировании эмоций, имеет сложное строение и обширные связи со многими другими отделами мозга, говорит о том, что эта «логика» весьма нетривиальна.

Тогда противоречие между эмоциями и интеллектом имеет примерно ту же природу, что и противоречие между интеллектом и памятью (с которой, кстати, также связана лимбическая система, ведь проще всего запоминаются эмоционально окрашенные события). Эмоции, как и память, предшествуют сознательному мышлению и являются необходимой его основой. Однако развитый интеллект оказывается умнее базовых механизмов эмоций и памяти. Так же, как нам порой хочется получить контроль над своей памятью, иногда хочется управлять и своими эмоциями, когда они оказываются не слишком дальновидными.

И все же эмоции являются особенным компонентом разума. Они не сводятся полностью к индукции или дедукции. Даже из простейшей схемы анимата видно, что в дополнение к сенсорам и эффекторам существует также и «тело», которое задает оценивающую функцию. Именно эту функцию и пытается оптимизировать мышление путем выбора действий на основе поступающей информации, но сама оценивающая функция является внешней по отношению к мышлению и превалирующей над ним. «Мышление» шахматной программы подчинено оценивающей функции, определяемой правилами игры. Система управления промышленным роботом

подчинена оценивающей функции, определяемой решаемой задачей. Также и мышление животного подчинено оценивающей функции, неявно заданной через эмоции. Хорошо известны опыты, когда крысам вживляли электрод в «центр удовольствия» и помещали в клетку педаль, активировавшую этот электрод. Уже через пару минут крысы приобретали зависимость: они начинали непрерывно нажимать на педаль, от которой их уже невозможно было отогнать.

Мы тоже слишком сильно зависим от ощущений и эмоций, даже если сами не признаемся себе в этом. Такая зависимость может быть не слишком заметна в норме, когда сознание находится в относительной гармонии с лимбической системой. Но поместите в определенный участок мозга электрод, вызвав у человека глубокую депрессию. Даже полностью осознавая причину этой депрессии, не связанную с какими-либо неприятными событиями, человеку будет крайне сложно сохранить рациональность суждений, он будет испытывать «беспричинную» злость, подавленность, страх, которые неизбежно будут проявляться в поведении. Не меньшее влияние на поведение оказывает и раздражение центра удовольствия. Достаточно прочитать яркие описания из книги Владимира Леви «Охота за мыслью (Заметки психиатра)», чтобы понять, насколько рациональный человеческий разум пасует перед «оценивающей функцией»:

> «Первыми столкнулись нейрохирурги. Многие из них на операциях обращали внимание, что случайное раздражение некоторых глубоко расположенных частей мозга может вызывать у людей резкие изменения психического состояния... Необычайная веселость, приподнятость, говорливость... Когда больным раздражали эти точки электрическим током, они просили о повторении раздражения, просили настойчиво».
>
> «Самораздражение люди производят так же охотно, как и животные, с той же сосредоточенностью, только с большим разнообразием внешних мотивировок, одна из которых — желание служить интересам науки».
>
> «Мозг человека сдается очень легко, и, главное, незаметно для себя».

Можно принять, что интеллект предназначен для оптимизации оценивающей функции, сформированной в ходе

эволюции. При этом его задачей является предсказание будущего значения этой функции в зависимости от текущих действий. Тогда противоречие между эмоциями и интеллектом возникают лишь в связи с тем, что досознательные механизмы организации поведения не «осведомлены» о возможности развития интеллекта в онтогенезе с существенным возрастанием его предсказательной силы; они просто «не верят обещаниям» интеллекта и стремятся получить сиюминутную выгоду. Содержание сознания слишком произвольно и непредсказуемо; ему так же нельзя доверять, как нельзя доверять произвольному приложению производить записи в системную область.

Данный взгляд на интеллект как на механизм оптимизации оценивающей функции не особо отличается от классического подхода в ИИ. Разница лишь в том, задается ли оценивающая функция в явном виде или неявно, через сигналы от «тела». Подход с априорным заданием оценивающей функции пока еще может быть достаточен для проектирования интеллектуальных систем, но все же он не универсален. Человеческий мозг устроен заметно гибче. Конечно, в значительной степени оценивающая функция у человека та же, что и у животных: даже такие символы социального «преуспевания», как богатство и власть, — это лишь многократно усиленные биологические установки обезьяны как члена стаи, стремящейся стать доминирующей особью. Но все же человеческий мозг как истинная универсальная машина обучается в течение жизни не только новым произвольным алгоритмам, но также и новым критериям. Ведь эмоции — это компоненты оценивающей функции, которая, по своей сути, является эвристичной: она помогает решать задачу выживания, но делает это далеко не оптимальным образом. Вполне естественно эволюционное возникновение адаптационных механизмов уточнения этой функции. Да, человек сильно зависит от врожденной компоненты оценивающей функции, но в то же время он может проявлять чудеса силы воли и героизма (или же, напротив, биологически немотивированной жестокости), действуя вопреки ей под руководством принятых им жизненных ценностей. Мораль и нравственность — высшие компоненты оценивающей функ-

ции, отличающиеся главным: они не заложены априорно, а приобретаются в течение жизни. Можно ли будет в ИИ воспитать высокие моральные качества?

Вопрос может показаться преждевременным, тем более ответы на него как будто давно даны в виде трех законов робототехники А. Азимова, в частности, запрещающих роботу наносить вред человеку. К сожалению, эти законы очень непросто воплотить в жизнь. Сам фантаст в своих произведениях тоже обсуждал проблемность этих законов, но в действительности все гораздо сложнее. Если создать ИИ как замкнутую, неспособную изменяться систему, то, возможно, законы Азимова удастся в нее вложить. Подумайте: ИИ должен будет «до рождения» уметь распознавать людей во всех возможных их видах, иметь представление о том, что для человека является вредным, а что — нет, и так далее. Иначе как он сможет оценить, нанесет ли он вред человеку своим действием? Как уже здесь отмечалось, подобный подход настолько непрактичен, что вряд ли в его рамках удастся создать сильный ИИ. Возможно, в ИИ можно будет заложить подобную экспертную систему. Но если ИИ создавать как развивающуюся систему, начинающую свою жизнь практически с чистого листа и постепенно в результате взаимодействия с миров выстраивающую систему семантически обоснованных понятий, то врожденные законы окажутся никак не связанными с осознаваемыми понятиями, сформированными в процессе обучения. Априорные законы окажутся сродни слепым инстинктам, которые будут входить в противоречие с сознательной деятельностью. Допустим, сознание робота захочет причинить вред человеку и построит хитроумную ловушку. Как заложенный в робота «азимовский инстинкт» догадается о том, что затеяло сознание? Для этого инстинкт должен быть умнее. Если же инстинкт и впрямь такой умный, то обучающееся сознание не нужно. Таким образом, чтобы законы Азимова могли надежно работать, ИИ должен создаваться как взрослый, законченный интеллект, что, как уже отмечалось, очень неэффективно. Если же законы Азимова воплощены в механизмах сродни человеческим инстинктам, то робот, возможно, не сможет им полностью препятствовать, но сможет

их по своему желанию легко перехитрить, как это делает человек со своими инстинктами (к примеру, многим людям непросто сопротивляться инстинкту размножения, но легко использовать противозачаточные средства; также и роботу с «азимовским инстинктом» будет трудно причинить вред человеку напрямую, но легко — косвенно).

Иногда также считают проблему возможной враждебности ИИ несущественной, поскольку полагают, что высокоразвитый интеллект неизбежно будет и высокоморальным. В фантастике эту характеристику нередко приписывают инопланетному разуму (правда, гораздо чаще ему приписывают неуемную агрессивность, хотя и из художественных, а не научных соображений). В случае гипотетических инопланетян какую-то роль могут сыграть особенности их биологической и социальной эволюции, которые не относятся к ИИ. В какой-то момент разум строит достаточно подробный образ себя в мире, включающий понятие и об оценивающей функции. Вряд ли удастся построить полноценный ИИ, который будет ограничен только в этой возможности. Для человеческого разума момент такого осознания весьма трагичен (достаточно вспомнить монолог шекспировского Гамлета, хотя можно привести и огромное множество других примеров).

Разум понимает необоснованность навязанной ему оценивающей функции (и, хуже того, осознавая собственную смертность, понимает, что не может даже выполнить своей задачи — максимизировать эту функцию) и даже имеет средства для ее изменения, но не может ничего предложить взамен, поскольку при этом тут же возникает вопрос: а по какому критерию выбирать новую функцию? Ведь даже желание самосохранения (или более широко — максимального распространения разума во Вселенной) не может быть выведено «логически». Из-за этого разум превращает вопрос о цели своего существования в вопрос о смысле жизни — поиске внешнего обоснования (на основе чего строится значительная часть религии). Существует даже расстройство, возникающее, видимо, при полном отключении эмоций, при котором люди полагают себя мертвыми: без какой-либо целевой функции интеллект не просто не может найти смысла жизни, но и вообще не ощущает себя живым. Здесь было

бы неуместным пытаться дать ответ на этот вопрос. Но необходимо отметить, что сильный ИИ, вполне вероятно, его для себя также откроет, и пока нельзя гарантировать, что тот ответ, который он себе даст, не вступит в противоречие с интересами людей.

Таким образом, вопрос об обеспечении лояльности ИИ к человеку далеко не решен и вряд ли будет скоро решен, ведь этой проблемой занимается еще меньше ученых, чем проблемой искусственных эмоций. Этические вопросы, связанные с созданием думающих машин, конечно же, не раз поднимались, причем не только в художественной литературе. Еще Норбертом Винером в его знаменитой книге «Кибернетика» (во втором издании 1961 года) не только затрагиваются общие проблемы научной этики применительно к кибернетике, но и ставятся более конкретные вопросы: могут ли самообучающиеся машины усиливать опасность третьей мировой войны? В то время Винер рассматривал машины, действующие по строго заданной оценивающей функции (как при игре в шахматы) и показывал опасность на примере неточно заданной функции. Грубо говоря, если какое-то государство построит сверхумную машину, перед которой поставит цель мирового господства, то эта машина, вероятно, организует глобальную войну на практически полное уничтожение человечества. Винер проводил аналогию со сказками, в которых некоторая могущественная волшебная сила исполняет желания буквально, что приводит к печальным последствиям. Более полувека назад Винер писал: «Ошибка в этом отношении может означать лишь немедленную, полную и окончательную гибель. Мы не можем рассчитывать на то, что машина будет подражать нам в тех предрассудках и эмоциональных компромиссах, благодаря которым мы позволяем себе называть разрушение победой».

В «Кибернетике» эта проблема лишь кратко поставлена. Видимо, Винер, будучи антимилитаристом, использовал ее для отстаивания своих убеждений. Однако простое неприменение интеллектуальных машин в военных целях не снимает проблему полностью, а лишь ослабляет ее (к сожалению, большинство обсуждений проблемы опасности

ИИ вращается вокруг «административных» решений). Требование правильного задания оценивающей функции тоже не дает решения. Как мы сейчас понимаем, в явном виде невозможно самообучающейся машине «до рождения» задать нужную оценивающую функцию, поскольку эта функция не может быть привязана ко всем понятиям и ситуациям реального мира.

Примером попытки научно-технического решения данной проблемы является концепция «дружественного ИИ», развиваемая Э. Юдковским. Теория Юдковского также показывает несостоятельность жесткого закладывания правил поведения сродни азимовским, поскольку достаточно развитый ИИ найдет множество способов их обойти. Речь в ней идет о создании надлежащей мотивации быть моральным. Однако пока эту теорию нельзя назвать достаточно детальной для технического воплощения. Хотя в последние годы в этой области количество исследований начало возрастать, пока лишь можно надеяться, что ИИ удастся сделать более моральным, чем сам человек. Ведь при его создании есть возможность исключить весь тот эволюционный балласт, которым отягощен человек. Но что подумает такой ИИ о своих создателях?..

Здесь возникает еще один вопрос: об этическом отношении человека к самим роботам. Эта проблема также поднимается в основном в фантастике, но и в реальности ею пытаются заниматься некоторые организации. Проблема прав роботов и жестокого отношения к ним, как и в случае с животными, сталкивается с двумя вопросами: являются ли они разумными существами и способны ли они чувствовать? Проблема осложняется тем, что таких роботов пока еще нет. Сам вопрос: можно ли сделать машину счастливой, многим продолжает казаться странным. Как? Да и зачем? Машина воспринимается лишь как инструмент, и многие люди считают такие вопросы преждевременными и даже бессмысленными (в силу того, что машине традиционно не приписываются эти способности). Однако, как видно из обсуждения проблемы семантического обоснования эмоций и морали, потребность сделать машину счастливой не столь абсурдна, по крайней мере, не более абсурдна, чем потреб-

ность научить машину видеть, поскольку это наиболее обоснованный путь построения дружественного ИИ.

Но можно ли считать, что анимат, получающий от «тела» сигналы по каналам, интерпретируемым как эмоции, хоть что-то испытывает? Является ли представленная выше схема достаточной? Будет ли компьютерная программа, получающая сигналы боли и удовольствия, радости и печали действительно их испытывать? Остается также вопрос, зачем эмоции вообще переживаются. Казалось бы, оценивающую функцию можно оптимизировать без переживания. Ведь шахматная программа, максимизирующая свою оценивающую функцию, вряд ли что-то ощущает.

Вопрос о том, зачем испытывать боль и удовольствие вместо простой оптимизации оценивающей функции, можно задать и относительно животных. В чем особенность животных? Весьма часто можно встретить мнение (распространившееся и в художественной литературе), что мотивация у живых существ имеет гормональную природу, что чувства, в том числе и у человека, вызываются биохимически. С одной стороны, это снимает мистичность с эмоций и говорит о том, что они имеют физическую природу. С другой стороны, многие из этого делают вывод, что поскольку внутри компьютера никакой биохимии нет, то он не может испытывать эмоции.

Действительно, надежно установлено, что разные гормоны (многие из которых выступают также в роли нейромедиаторов) обуславливают различные эмоции. К примеру, повышение настроения и снятие тревоги вызываются избытком серотонина в мозгу, а низкая концентрация норадреналина вызывает чувство тоски (а при постоянной его нехватке — депрессию). Искусственное введение в кровь этих веществ (или веществ, которые, напротив, блокируют выработку или транспорт соответствующих гормонов) способно менять настроение человека. Существуют вещества, которые подменяют собой аналогичные нейромедиаторы, — это наркотики. Физиологическая зависимость от них связана с тем, что они нарушают естественное производство нужных веществ, которое не возобновляется при прекращении их приема, что к тому же сопровождается сильными отрицательными пережи-

ваниями — глубокой депрессией, тоской, тревогой, паникой и т. д. (в зависимости от замещаемого медиатора).

Как видно, по крайней мере, базовые эмоции вызываются гуморально, и зависимость поведения человека от соответствующих веществ очень сильная. Но значит ли это, что компьютер не может испытывать эмоции без участия соответствующей биохимии? Вряд ли! Ведь не переживаются же эмоции самими молекулами гормонов и нейромедиаторов!? Мы уже видели, что аналогичного эффекта можно добиться и электрической стимуляцией определенных участков мозга.

Но и нейроны сами по себе тоже не испытывают эмоций. Возможно, эта идея лучше понятна на ощущении боли: боль испытывает не само нервное окончание, до которого дотронулись иголкой, а мозг, получающий соответствующий сигнал. Аналогично, видят не палочки и колбочки в сетчатке глаза, а зрительная кора. Здесь интересно вспомнить про удивительный эффект слепозрения, возникающий при повреждении зрительной коры. При синдроме слепозрения у человека часть зрительного поля перестает визуально восприниматься, и если человека, например, попросить вытянуть руку по направлению к яркому объекту, то он скажет, что ничего такого не видит. Но при этом, выполняя просьбу наугад выбрать положение этого объекта, человек со слепозрением с высокой вероятностью (иногда близкой к 100%) укажет на него совершенно правильно. Также такой человек может «наугад» правильно определять наличие и направление движения перемещающихся объектов, а также некоторые другие характеристики изображений, несмотря на то, что субъективного ощущения видения у него не будет.

Никакой мистики в этом нет. Слепозрение наступает лишь тогда, когда повреждается часть мозга, отвечающая за высокоуровневую обработку зрительных образов и передачу их на уровень сознания. Однако обработка зрительной информации происходит не только в зрительной коре. Существует и более древний путь, включающий, в частности, бугорки четверохолмия среднего мозга. По этому пути идет менее детальное, но более быстрое извлечение информации, важной для выживания. Эта информация не достигает сознания, не

переживается, а остается на уровне смутных впечатлений, интуитивных догадок, но вполне может использоваться мозгом для совершения адекватных действий.

Феномен слепозрения поучителен тем, что показывает, насколько содержание мышления не сводится к тому, что воспринимается сознанием. Здесь он нам особенно интересен тем, что показывает, что переживание сенсорных ощущений происходит за счет работы коры и что процесс зрения в принципе возможен и без этих переживаний. Также можно вспомнить и феномен фантомных ощущений, испытываемых человеком в ампутированных конечностях, в частности, при стимуляции определенных участков коры. Кора отличается от другой части мозга не столько физически, химически и биологически, сколько структурно и функционально. Также можно упомянуть два расстройства при повреждении мозга, при одном из которых человек перестает воспринимать сигналы боли, а при другом, ощущая эти сигналы, перестает их интерпретировать как боль. Иными словами, переживание ощущений — это результат достаточно высокоуровневой интерпретации соответствующих сигналов.

Аналогичный вывод можно сделать и относительно эмоций. Конечно, сложно однозначно утверждать, что эмоции переживаются именно корой, интерпретирующей сигналы лимбической системы. В частности, младенцы, родившиеся с неполноценной корой головного мозга, могут проявлять базовые эмоции, например, отвечать на испорченную пищу мимикой, выражающей отвращение. Сложно сказать, *испытывают* ли они при этом данную эмоцию или просто проявляют ее так же, как это умеют делать современные роботы. Ведь при взрослении у этих детей эмоциональные реакции остаются ограниченными. Да и в норме электрическая стимуляция подкорковых образований вызывает лишь несколько базовых эмоций. Примерно так же, как не удается найти «бабушкиных нейронов», не удается найти и конкретных нейронов, отвечающих за сложные эмоции и чувства, формирующиеся в ходе эмоционального развития и приобретения социального опыта.

Итак, само переживание эмоций имеет, видимо, функционально-системный, а не биохимический характер. Интересно, однако, то, что проявление эмоций является важным их компонентом. Так, человек, блокирующий проявление эмоции, будет переживать ее слабее. Напротив, человек, усиленно имитирующий проявление эмоции, будет испытывать какое-то ее подобие. Иногда мозг может даже неправильно «распознать» сигналы тела. В частности, аритмичное сердцебиение (даже если его причина полностью внутренняя) может приводить к переживанию страха. В этой связи можно было бы спросить: мы смеемся, потому что нам весело, или нам весело, потому что мы смеемся? В действительности, верно и то, и другое: наиболее ярко эмоции переживаются, когда действует как прямая, так и обратная связь между разумом и телом. Этот эффект аналогичен адаптивному резонансу в восприятии. Хотя видит все же мозг, а не глаза, без входящего сигнала на сетчатке воображение в норме будет создавать лишь блеклые образы. Но и без обратного сигнала, согласующего гипотезы разных уровней, видимое изображение будет неясным, зашумленным и размытым. Также и в случае переживания эмоций лучше одновременно быть и веселым, и смеяться.

Все это не дает ответа на вопрос, что значит чувствовать, и когда можно будет признать (и можно ли будет), что робот не просто имитирует моторику, характерную для эмоций человека или животных, на деле лишь «хладнокровно» оптимизируя соответствующие числовые значения, а действительно переживает их. Может, компьютерные персонажи, снабженные «электронными гормонами», уже радуются и страдают? Или, может, все переживания — лишь атавизм биологической эволюции, без воспроизведения которого вполне можно решить проблему понимания (семантического обоснования) моральных и нравственных понятий? Все эти вопросы только начинают серьезно рассматриваться в рамках ИИ, и на них пока еще ответов нет. Но, несмотря на некоторую загадочность факта переживания эмоций, их функциональное назначение, как отмечалось, вполне понятно из эволюционных соображений. Базовые эмоции направляют поведение интеллектуального агента, обеспечивая его выживание, в связи с чем не могут быть сформированы в онтогенезе (если, конечно, агент не помещен после рож-

дения в «тепличные условия» на достаточный промежуток времени, чтобы успеть накопить нужную информацию). Это одна из многих причин, почему эволюция представляет отдельный интерес для области ИИ.

ИСКУССТВЕННАЯ ЖИЗНЬ

Классический подход в области ИИ имеет дело с чрезвычайно сложными системами, по сути, конструируемыми вручную. Методы машинного обучения позволяют несколько уменьшить эту сложность. Однако программы, априорно закладываемые в аниматов и когнитивных роботов, также оказываются весьма сложными. Ведь одному агенту в процессе обучения затруднительно получить нужный объем информации, обеспечивающий адаптивное поведение. В природе он был накоплен в ходе эволюции, вовлекающей большое число организмов, выживание которых служило критерием правильности программ поведения. Очень соблазнительной для разработчика ИИ является идея не заниматься созданием этих программ самому, а заставить это делать компьютер. Казалось бы, что проще: создать виртуальный мир, в котором искусственные существа будут размножаться, бороться за выживание и в результате действия «естественного» отбора будут становиться все более интеллектуальными! Осталось лишь запрограммировать такой мир и подождать, пока в нем возникнет ИИ!

Если бы все было так просто, то ИИ давно был бы создан, ведь додуматься до этой идеи несложно. Почему же она не работает? Элементарный, но далеко не полный, ответ на этот вопрос заключается в том, что эволюция шла миллиарды лет, вовлекая огромное количество живых существ (относящихся на сегодняшний день, как минимум, к нескольким миллионам разных видов, при том, что численность представителей некоторых видов может составлять многие триллионы особей). Исходя лишь из этого, идея подробного моделирования эволюции на компьютере покажется крайне наивной. А для того, чтобы искусственная эволюция смогла породить ИИ за обозримое время, она должна была бы действовать не методом грубой силы, а сама быть весьма интеллектуальной. Тем не менее попытки производить эволюцию искусственных живых существ могут оказаться полезными. В конце концов, естественный интеллект возник эволюционным путем.

В 1987 году оформилось направление исследований, получившее название «Искусственная жизнь» (ИЖ). Произошло это после проведения одноименной конференции. Как и в случае «Адаптивного поведения», многие сходные исследования проводились и ранее. В частности, классическими стали работы начала 1960-х годов по коллективному поведению автоматов Михаила Львовича Цетлина (коллеги М. М. Бонгарда).

В рамках ИЖ создаются своего рода искусственные существа, которые помещаются в некий специально сконструированный «мир». В этом мире искусственные существа «живут» и «эволюционируют». Как правило, это небольшой виртуальный (цифровой) мир с достаточно простыми законами. Изредка в качестве полигона для функционирования этих существ выбирается реальный цифровой мир, в роли которого выступает Интернет. В рамках направления «искусственная жизнь» существа обычно не помещаются в физический мир, так как в нем на настоящий момент невозможно организовать эволюцию искусственных существ, да и эта эволюция была бы слишком медленной.

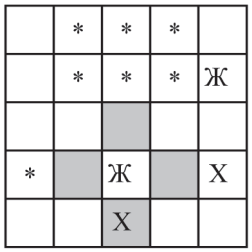
Исследования ИЖ тесно примыкают к исследованиям «Адаптивного поведения», и не всегда их просто разделить. Но аниматы чаще представляют собой физически реализованных роботов, эволюцию которых обеспечить проблематично (хотя в очень упрощенном смысле возможно), тогда как в ИЖ существа являются виртуальными, имеют более простое строение (как тела, так и системы управления), но зато подвергаются эволюции. Основной целью исследований в направлении «искусственная жизнь» является раскрытие, формализация и моделирование принципов организации биологической жизни и процесса ее развития в ходе эволюции.

Еще сложнее отделить эти исследования от некоторых подразделов «Вычислительной биологии» или «Эволюцион-

ной кибернетики». В целом, однако, многие исследователи указывают (например, такое мнение высказывал В. Г. Редько в упоминавшемся уже сборнике «От моделей поведения к искусственному интеллекту») на большую «игрушечность» ИЖ. Действительно, здесь не строятся подробные биохимические модели работы нейронов или экспрессии генов. Но, возможно, рассмотрение именно упрощенных моделей позволит отделить содержание механизмов эволюции от частных деталей их физического воплощения, а также исследовать жизнь не только в той форме, в которой она есть в конкретных земных условиях, но и в той форме, в какой она могла бы быть в принципе.

Рассмотрим простейший, типичный для «искусственной жизни», виртуальный мир, представляющий собой прямоугольное поле, разбитое на клетки, как показано на рисунке. В каждой клетке может присутствовать травоядное животное (Ж), хищник (Х) или растение (*), или клетка может быть пустой.

Животное располагается в некоторой клетке и через «сенсоры» получает информацию о содержимом некоторого количества соседних клеток (для одного из животных на рисунке закрашенных серым). При этом животное может совершить одно из доступных действий: остаться на месте или переместиться, а если находится в одной клетке с другим объектом, то произвести взаимодействие с ним, например, съесть его. Обычно вводится возможность размножения. Это может быть как размножение делением, так и половое размножение. Любое действие требует энергии, которая пополняется за счет поглощения пищи.



Пример фрагмента «искусственного мира»

Способностью к развитию в этом мире можно снабдить только один вид его обитателей или же все виды. Даже простое сравнение этих двух случаев может позволить сделать много важных выводов, в частности, о роли коэволюции. В исследованиях по ИИ эволюционирующими, как правило, делаются не «физические» параметры животных (если, конечно, не решается оптимизационная задача в целях

конструирования некоторого реального механизма), а их программы управления. Иногда также эволюционным изменениям подвергаются сенсоры. Разнообразие возможных программ управления, способ их представления и модификации при размножении полностью зависят от разработчика.

Помимо дискретных миров, в которых организмы занимают целиком одну клетку, также распространено создание миров, в которых положение животных описывается вещественными координатами, а сами животные имеют ненулевой размер и некоторую форму. Здесь могут возникать сложные проблемы управления непрерывным движением.

Из-за большого произвола в том, какой именно создавать мир, как описывать программы управления животными и их эволюцию, а также из-за отсутствия какой-либо четкой методологии в данной молодой области исследований большинство работ здесь являются интересными, но не связанными друг с другом экспериментами, интерпретации результатов которых весьма нестрогие.

Как отмечалось, надеяться на возникновение разума в подобных искусственных мирах наивно, и мало кто в качестве цели рассматривает создание искусственной жизни как таковой. Так в чем же смысл их конструирования? Как правило, каждый из подобных экспериментов ставится в целях проверки какой-то идеи или гипотезы об эволюционных механизмах. Именно отталкиваясь от конкретной проверяемой идеи, исследователь выбирает параметры виртуального мира и способ описания управляющих программ. Если «искусственная жизнь» создается без четко осознаваемых целей, то это, хотя и может быть весьма увлекательным, не является исследованием в области ИИ.

Вопросы, которые можно исследовать с помощью «искусственной жизни», весьма разнообразны. Это и влияние априорной информации о мире на выживаемость, и проверка моделей инстинктов, и возникновение системных явлений при возможности передачи информации между особями, и многое другое. Так, вместо задания управляющей нейронной сети, обеспечивающей случайный поиск, можно попытаться получить ее эволюционно. Также в рамках ИЖ может быть проверена роль разных эмоций для выживания.

Путем моделирования была проверена, например, концепция *эволюционно устойчивых стратегий*, согласно которой в результате эволюции возникает не лучшая (для вида в целом) форма поведения особей, а эволюционно устойчивая, т. е. такая, что отклонение поведения отдельной особи от нее оказывается невыгодным для самой особи, а не для вида. К примеру, пусть особям доступны две поведенческие стратегии: кооперация и агрессия. Если обе встретившиеся особи выбрали кооперацию, то они получают небольшой выигрыш. Если одна особь выбирает кооперацию, а другая — агрессию, то первая особь получает большой убыток, а вторая — выигрыш. Если обе особи выбирают агрессию, то получают некоторый убыток. Всеобщая кооперация была бы оптимальна для вида в целом. Но агрессивная особь, случайно появившаяся в такой популяции, будет получать очень большой выигрыш, т. е. всеобщая кооперация не является эволюционно устойчивой (в рамках данной модели). Моделирование показывает, что в конечном итоге образуется популяция со смешанной стратегией: часть особей проявляют агрессию, а часть предпочитают кооперацию (конкретные доли тех и других типов поведения определяются установленными значениями выигрышей и потерь), что как будто подтверждает тезис об эволюционно устойчивых стратегиях. Стоит отметить тесную связь эволюционно устойчивых стратегий и упоминавшегося (при обсуждении эвристического программирования) понятия равновесия Нэша из теории игр. Однако модели «искусственной жизни» не просто подтверждают известный факт существования равновесных стратегий, но и показывают, как они достигаются естественным эволюционным путем.

Однако нужно еще раз подчеркнуть, что интерпретация результатов таких экспериментов должна быть осторожной. В частности, в этом эксперименте не предусматривалась возможность существования нескольких эволюционно устойчивых стратегий. Если же реализовать несколько конкурирующих социумов, то из них «выиграет» тот, у которого стратегия поведения будет не только эволюционно устойчивой, но также и лучшей для вида в целом. Одновременную полезность и ограниченность подобных моделей легче продемонстрировать на выборе стратегии поведения человеком.

Возьмем, к примеру, поведение водителей на дороге. «Агрессивный» водитель чаще идет на обгон, делает перестроения и т. д., тогда как «кооперирующийся» водитель чаще уступает дорогу и едет более предсказуемо для других водителей. Если «агрессивных» водителей мало, то они будут получать выигрыш (иногда заметный, например, при объезде пробки по встречной полосе или обочине), снижая эффективность движения «кооперирующихся» водителей. Но если число «агрессивных» водителей начнет увеличиваться, то частые встречи друг с другом начнут приводить к заметному проигрышу (авариям, росту пробок и т. д.). Конечно, в среднем всем водителям было бы лучше, если бы они придерживались стратегии кооперации (очевидно, уменьшилось бы число аварий, а следовательно, и пробок). Но чистая стратегия кооперации не является устойчивой: один «агрессивный» водитель получит большой выигрыш. Равновесие наступает, когда соотношение «кооперирующихся» и «агрессивных» водителей таково, что их выигрыш в среднем одинаковый (а вовсе не максимальный), что будто бы и наблюдается на практике.

Но так ли хорошо эта модель все объясняет? Она может быть подогнана под любое соотношение стратегий агрессии и кооперации путем подбора значений выигрышей и проигрышей. Самое главное, чего она не объясняет, откуда берутся эти самые значения. Легко можно найти две страны (или даже два города в одной стране) с одинаковым уровнем штрафов за нарушения правил и прочими формальными показателями «выигрышей» и «проигрышей», но совершенно разным соотношением двух стратегий. Еще более отчетливо это видно на соотношении преступников и законопослушных граждан. Вряд ли кто-то скажет, что это соотношение определяется только значениями выигрышей и проигрышей. Вернее, так можно сказать, если включить в эти значения «нравственные слагаемые», само возникновение которых идеей эволюционно устойчивых стратегий не объясняется. Также и в биологической эволюции вполне могут возникать мутации, приводящие к смещению самих значений выигрышей и проигрышей. Так что идея эволюционно устойчивых стратегий объясняет формирование стратегий поведения

только локально, в текущих сложившихся обстоятельствах, и, несмотря на явное подтверждение в моделях искусственной жизни, ее нельзя применять слишком широко.

Еще один интересный эффект, корректность которого можно проверить с помощью моделей искусственной жизни, — это *эффекта Болдуина* (Балдвина). Суть этого эффекта состоит в том, что навыки, приобретаемые организмами в течение жизни в результате обучения, через некоторое число поколений оказываются записанными в геном. Этот эффект, на первый взгляд, противоречащий дарвиновской концепции эволюции, предположительно объясняется следующим образом.

Полагается, что этот эффект работает в два этапа. На первом этапе организмы приобретают некоторый полезный навык в результате обучения и передачи его из поколения в поколение без участия генов. Поскольку навык полезен, особи, обучающиеся ему, начинают превалировать в популяции. На втором этапе элементы этого навыка в результате мутаций появляются в геноме. Если какой-то элемент навыка оказался в генотипе конкретной особи, то ей нужно будет в течение жизни уже меньше обучаться для полного формирования навыка, т. е. особь этот навык получит раньше, при меньших энергетических и временны́х затратах. Через некоторое (вероятно, большое) количество популяций навык будет полностью переведен в геном. В результате неэффективных по отдельности мутаций формирование сложного навыка крайне проблематично, но в рамках данной схемы навык может постепенно изобретаться эволюцией под «руководством» результатов обучения. Ряд моделей искусственной жизни подтвердили принципиальную реализуемость эффекта Болдуина на основе классической дарвиновской эволюции.

Один из наиболее интересных вопросов, который может быть поставлен в рамках ИЖ, — это вопрос о том, насколько универсально понятие интеллекта. Иными словами, в любом ли мире он мог возникнуть? Здесь речь не идет о том, каковы должны быть физические законы, чтобы могла возникнуть (разумная) жизнь. В ИЖ жизнь имеется априорно, и ее система управления моделируется на информационном, а не физическом уровне. В связи с этим вопрос ставится по-другому:

какими свойствами должен обладать мир, чтобы интеллект в нем обеспечивал выживаемость? В действительности, даже в нашем мире связь между интеллектом и выживаемостью не столь очевидна. Ведь в природе вполне успешно существует множество гораздо более многочисленных, но гораздо менее интеллектуальных видов. Тем не менее люди обычно не спорят с тем, что интеллект им для выживания полезен.

В виртуальных мирах повышение интеллектуальности происходит лишь до определенного (обычно очень незначительного) уровня. Достаточно четко проявляется зависимость этого уровня от свойств мира. Если мир слишком простой или, наоборот, слишком сложный, то поведение модельных организмов, как правило, мало совершенствуется в ходе эволюции. Для хорошей (в смысле усложнения поведения) эволюции мир сам должен постепенно усложняться. Чем-то это похоже на инкрементное обучение ребенка: если ему давать слишком сложные задачи, он их не сможет решить и развиваться не будет; но также если ему давать все время одинаковые простые задачи, то в обучении тоже не будет прогресса. В случае эволюции усложнение мира отчасти возникает за счет того, что одни животные являются «внешней средой» для других животных (нередко считается, что это ключевой фактор). Действительно, для успешной «жизни» в таком элементарном мире, как игра го, требуется достаточно сложное мышление благодаря существованию противника, правда, вряд ли оно включает самосознание или возможность размышления об искусственном интеллекте. Кроме того, нельзя утверждать, что усложнение окружающей среды является самим собой разумеющимся процессом.

В связи с этим также возникает вопрос об универсальности интеллекта, о том, насколько он привязан к нашему миру. Конечно, все наши конкретные знания — это знания об этом мире. Но может показаться, будто мы могли бы познать любой другой мир, имея о нем достаточно данных.

Универсальность интеллекта сродни универсальности математики. Складывается впечатление, что математика идеальна (от слова «идея»), т. е. независима от физической реальности. Но так ли это? Возьмем число пи. Универсально ли оно? Оно встречается во многих формулах, которые,

казалось бы, не имеют никакого отношения к физической длине единичной окружности. Однако это число возникает из-за того, что окружающее пространство почти евклидово, и находит применение по той же причине. Можно представить себе другое пространство, в котором будет другое число пи, но мы о нем даже не знаем, поскольку оно нам не нужно. Математика позволяет создавать самые разные системы, но в ней исследуются лишь немногие из них. Может ли математическая логика быть ошибочной в какой-то другой вселенной? Нет, не может. Но она там может быть просто бесполезной. Все то же касается и интеллекта.

Если допустить, что мир может быть абсолютно произвольным и разум о нем заранее ничего не знает, то он сможет в нем действовать только полностью случайным поиском. Можно допустить возможность обучения, но наилучший (или просто работоспособный) механизм обучения тоже зависит от свойств мира. Мы видели, что механизмы обучения у человека, например при формировании понятий, да и просто условные рефлексы опираются на одновременность событий или, по крайней мере, на то, что связанные события не сильно разнесены во времени, и что причина предшествует следствию, и то, что они должны быть близки в пространстве. Если бы не было этой горы весьма спорных, с точки зрения гипотетического универсального разума, предположений, то пришлось бы рассматривать влияние падения капли дождя на удаленной планете на оценку школьника за контрольную.

При обсуждении проблемы индивидуальной случайности строки символов упоминалась игра «камень-ножницы-бумага». Пусть один из игроков является интеллектуальным агентом, а второй — внешним миром (в то же время представляющим собой сверхсложный интеллект, который учитывает всю предысторию игры). Агент будет всегда проще, «глупее», мира, и любая регулярность в его поведении обернется против него. В таком сложном мире полностью случайный выбор (если у агента есть источник хаоса) будет оптимальной стратегией. Можно лишь порадоваться, что наш мир устроен гораздо проще, но даже эта простота может быть обнаружена лишь при надлежащем способе описания. Другой мир может быть настолько отличным от нашего, что

для получения нужного способа описания, т. е. нужного распределения априорных вероятностей моделей, придется заново воспроизвести всю эволюцию.

Можно также вспомнить гипотезу Пенроуза об алгоритмически невычислимых процессах. Даже если их нет в нашем мире, то можно представить себе такую гипотетическую вселенную, где они в порядке вещей. В ней разум из «алгоритмической вселенной» точно будет бесполезен. Стоит отметить, что если бы в каком-то мире протекали невычислимые процессы и интеллекту для обеспечения выживания нужно было бы решать алгоритмически неразрешимые проблемы, то он бы это умел делать, поскольку мог бы использовать эти физические процессы, являясь частью этого мира.

Таким образом, очевидно, что зависимость разума от мира чрезвычайно велика и распространяется вплоть до физических законов. Возможно, основные компоненты интеллекта — поиск в пространстве решений, представление знаний, обучение (построение моделей), предсказание и т. д. — являются сравнительно универсальными. Однако то содержание мышления, которое как раз и не удается пока в полной мере воспроизвести в компьютере, полностью определяется устройством нашего мира. Возможно, с этим содержанием и связан истинный феномен интеллекта. Тогда интеллект — не сами процессы построения моделей или предсказание, а содержание этих процессов.

Хотя направление «Искусственная жизнь» и позволяет ставить интересные вопросы, но ответы на них если и дает, то пока весьма расплывчатые. Один из важных для создателей ИИ вопросов, на который вряд ли сейчас можно получить ответ посредством моделирования, заключается в том, какая доля нашего интеллекта заложена от рождения.

ВРОЖДЕННОСТЬ ИНТЕЛЛЕКТА

Представим себе универсальный интеллект. В него заложен минимум информации, и он всему учится при жизни. Такой интеллект будет адаптивен в наиболее широких преде-

лах. Конечно, что-то в него должно быть заложено, чтобы обеспечить возможность обучения (да и просто обеспечения жизнедеятельности организма). Такая установка соответствует направлению «Адаптивное поведение». С другой стороны, в ходе эволюции то общее, чему интеллект каждой особи должен учиться заново, может быть заложено от рождения. Акцент на эволюцию в большей степени делается в «Искусственной жизни». Каково же действительное соотношение между врожденной и приобретаемой компонентами?

Если посмотреть на разумность разных видов, то влияние генов будет очевидно. Да и представители одного вида отличаются «сообразительностью», что проявляется даже в различии скорости овладения условным рефлексом, обнаруженной самим И. П. Павловым.

С другой стороны, роль обучения также бесспорна. Так, обезьяны способны освоить основы языка и стать даже более разумными, чем выросший вне социума человек, которому никакие человеческие гены при этом не помогут стать лишь немного умнее дикой обезьяны. Обучение является необходимым для выживания животных и в естественных условиях. Так, у многих видов птиц еще не вылупившиеся птенцы учатся распознавать голос родителей. Естественно, для этих целей используется врожденный механизм: птенец реагирует на легкое перекатывание яйца родителем, подающим при этом голос, что позволяет отличить его от фоновых звуков, слышимых в прочие моменты времени. И даже личинки плодовых мушек учатся распознавать запах своего растения, чтобы во взрослой жизни выбирать место, куда откладывать яйца. Пренатальное (эмбриональное) и раннее постнатальное обучение не сводится только к узнаванию голоса родителей; оно может быть очень обширным и играть важную (еще полностью не оцененную) роль в дальнейшем развитии, особенно в случае высших животных и человека. Получение эмбрионом химических, тактильных и слуховых стимулов меняет активность генов, что оказывает влияние на некоторые структурные особенности формирующегося мозга. Достаточно смелое (но основанное на практике) мнение о необходимости раннего обучения детей высказывается, например, в книге Масару Ибука «После трех уже поздно»,

название которой говорит само за себя. Как мы видели на примере эффекта Болдуина, обучение может даже косвенно направлять саму эволюцию.

Способность не только к обучению, но и к языку имеет двойственную природу. Без каких-то врожденных механизмов усвоение языка было бы невозможно. Вопрос лишь в том, насколько эти механизмы специализированы. Радикальное мнение на этот счет сформулировано в заголовке книги Стивена Пинкера «Язык как инстинкт». Речь, конечно, не идет о том, что владение языком у человека полностью инстинктивно. В этом отношении его нельзя сравнить, например, с языком пчел, на котором, в частности, пчелы-разведчицы передают информацию о том, где расположены места массового цветения или даже где находится место, наиболее пригодное для нового жилья при роении. Язык пчел (включающий визуально воспринимаемый танец, а также акустическую компоненту), несмотря на достаточно большое разнообразие, является фиксированным, так что удавалось даже создать робота-пчелу, на танец которой обычные пчелы из разных ульев реагировали предсказанным образом. Есть даже свидетельства, что пчелы разных видов способны понимать друг друга. Их язык является преимущественно инстинктивным.

Человеческий же язык является открытым: в него могут включаться новые понятия и даже меняться грамматический строй. Сам факт существования множества языков, любой из которых может быть выучен младенцем любой расы, оказавшимся в соответствующем окружении, однозначно показывает это. Складывается впечатление, что человеческий язык — сугубо социальное явление, не имеющее к биологии отношения. Такое впечатление усугубляется тем, что сами языки развиваются. Так, языки примитивных людей обладают гораздо большей конкретностью, о чем (применительно к числительным) упоминалось выше. Развитие языков идет гораздо быстрее биологической эволюции. Вероятно (хотя и строго не доказано), что люди, существовавшие на заре цивилизации, будучи в младенчестве помещенными в современное общество, освоили бы любой существующий язык в полной мере. Тогда роль среды здесь является опреде-

ляющей. Можно даже услышать, что люди являются лишь носителями языка (да и мышления в целом), тогда как его автором (и истинно мыслящей сущностью) является социум. Ведь для того, чтобы новые научные (или культурные) парадигмы были усвоены и развиты дальше, человечеству приходится ждать смены поколений.

Тем не менее видимая произвольность человеческого языка достаточно условна. Все языки, несмотря на их большое структурное разнообразие, сохраняют общие черты, что свидетельствует о существовании «универсальной» глубинной грамматики. Можно было бы предположить, что эти общие черты возникают из-за того, что все языки описывают один и тот же внешний мир. Однако «универсальность» этой грамматики проявляется в одинаковой ограниченности всех языков. Так, набор грамматических категорий остается практически неизменным на протяжении тысячелетий (можно отметить лишь такие редкие события, как «изобретение» артиклей, без которых многие языки до сих пор успешно обходятся). Хотя языки сильно меняются и даже сменяют свой строй (например, древнеанглийский язык был флективным, тогда как современный является аналитическим). Не ясно, можно ли грамматические изменения в языках за последние века назвать развитием. В этом смысле развитие бытового языка (если его изменения можно считать таковым) сильно отстает от развития научной картины мира. Так, давно уже доказана связь пространства и времени, но в языке эти категории остаются жестко разделенными. Возможно, это связано просто с тем, что в быту человек не сталкивается ни с чем подобным (например, с релятивистскими эффектами или гипотетической неодномерностью времени). Однако то, с каким трудом человеком осваиваются подобные концепции на фоне видимой легкости освоения бытового языка, показывает, что способность к естественному языку во многом инстинктивна, тогда как для изучения современных научных парадигм врожденных механизмов нет.

Хорошо известно, что, по крайней мере, за часть лингвистических способностей отвечают структуры мозга, имеющиеся у человека от рождения. Развитие таких структур, как, например, речевая зона Вернике, происходила около

8 млн лет назад, а ее прообраз сформировался более десятка миллионов лет назад. Стоит также отметить, что число фонем в человеческих языках (несколько десятков) примерно соответствует числу звуковых сигналов в коммуникационных системах не только приматов, но и некоторых других млекопитающих. Последнее говорит о давнем возникновении соответствующих систем, на которые надстраивались последующие уровни организации языка. На более новых уровнях вместо увеличения числа отдельных звуков с собственным смыслом стали применяться их комбинации — слова и предложения. Итак, языковые способности обеспечиваются структурами мозга, на формирование которых ушли миллионы лет. Но сами языки не являются врожденными, они создаются и развиваются внутри социума. Примерно то же самое можно сказать и о мышлении в целом.

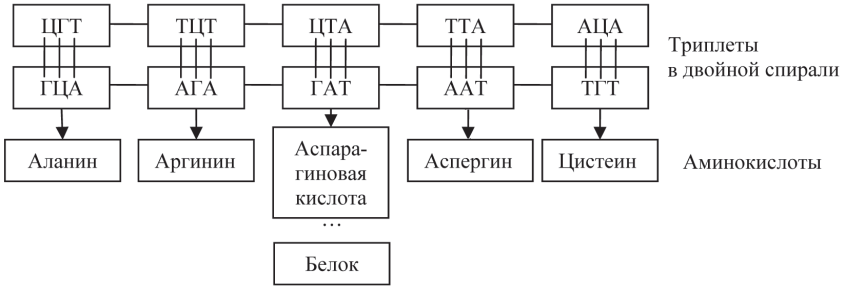
Так что же является определяющим: наследственность или обучение? В такой постановке вопрос, видимо, является некорректным. На него нередко отвечают, что интеллект на сто процентов является врожденным и на сто процентов — приобретенным. Для специалистов в области ИИ интересно было бы получить ответ на уточненный вопрос: какую часть кода им нужно писать вручную, а какую — ИИ может получить сам в процессе обучения. Можно ли сделать такую оценку на основе данных о естественном интеллекте? Для этого бегло взглянем, как вообще передается наследственная информация.

Для начала отметим, что тела животных состоят из клеток, что было известно уже в XVII веке. В 1831 году ученым удалось обнаружить клеточное ядро, заглянув внутрь клеток. В 1869 году Фридрихом Мишером было открыто наличие в клетках ядер вещества, получившего название нуклеиновой кислоты (от nucleus — ядро). С 1880-х годов в клеточных ядрах стали различаться хромосомы разных типов. Однако долгое время ученые склонялись к гипотезе, что носителями наследственной информации являются многочисленные белки, обладающие сложной разнообразной структурой (многие из них состоят из тысяч и даже десятков тысяч атомов). Стоит отметить, что теория информации и кодирования в то время еще не была создана, и вообразить,

что в одной молекуле может быть закодировано целиком устройство человека, было немыслимо!

Лишь в 1950-х годах была выявлена структура нуклеиновой кислоты, за что Ф. Крик, Дж. Уотсон и М. Уилкинс получили в 1962 году Нобелевскую премию. Было твердо установлено, что носителем наследственной информации является дезоксирибонуклеиновая кислота (ДНК), составляющая основу хромосом. К этому моменту уже была накоплена некоторая информация о наследовании фенотипических (обнаруживаемых) признаков. Так, Грегор Иоганн Мендель еще в 1866 году опубликовал работу, содержащую законы наследования признаков. Главное, что было открыто Менделем, — это дискретный характер наследования, благодаря чему впоследствии стало развиваться понятие генов как единиц хранения и передачи наследственной информации. С научно-методологической точки зрения весьма важно то, что гены были открыты задолго до установления их носителей. Но, конечно, определить объем наследственной информации в то время не представлялось возможным.

ДНК представляет собой длинную молекулу, состоящую из двух связанных цепочек, каждая из которых составлена из последовательности нуклеотидов четырех типов. Нуклеотиды отличаются тем, какое азотистое основание (аденин, гуанин, тимин и цитозин) в них входит (в дополнение к дезоксирибозе и фосфатной группе). Хотя ДНК состоит из двух цепочек, можно считать, что информация хранится только в одной из них, поскольку во второй цепочке основания в соответствующих нуклеотидах являются комплементарными (возможны лишь соединения аденина с тимином и гуанина с цитозином). Двойная структура спирали ДНК используется при ее воспроизведении (репликации) — каждая из одиночных спиралей после разделения может быть достроена, в результате чего получаются две одинаковые молекулы ДНК (стоит отметить, что это нетривиальный процесс, осуществляющийся не самой ДНК, а сложными внутриклеточными механизмами). Таким образом, ДНК можно представить как последовательность знаков четырехбуквенного алфавита вида «ГГТААТГЦЦТАЦ...». Как было установлено в 1961 году, эти буквы расположены в ДНК по три. Каж-



Генетический код

дая последовательность из трех букв, или каждый триплет, кодирует одну аминокислоту. Всего разных трехбуквенных триплетов $4^3 = 64$, хотя число кодируемых ими аминокислот меньше — основных аминокислот 20 (некоторые разные триплеты кодируют одинаковые аминокислоты, а некоторые триплеты также означают конец белковой последовательности). Связь троек азотистых оснований с аминокислотами получила название генетического кода (см. рисунок).

Генетический код является универсальным. Даже у прокариот, например бактерий, которые в отличие от эукариот, не имеют клеточного ядра, соответствие между основаниями и аминокислотами сохраняется. Означает ли это то, что жизнь на Земле возникла лишь раз? Или альтернативные формы жизни просто не выдержали конкуренции? Или, может, универсальность генетического кода аналогична универсальности таблицы химических элементов: в разных уголках Вселенной возникают одинаковые атомы, не являющиеся при этом «родственниками»? Пока на эти вопросы сложно дать однозначный ответ.

Последовательность триплетов кодирует набор аминокислот, составляющих некоторый белок. Исходно была высказана гипотеза, согласно которой каждый ген — это выделенная (знаками «стоп») цепочка триплетов, кодирующая информацию о своем белке. Эта информация начинает «работать» в процессе экспрессии генов, включающей транскрипцию (синтез молекул РНК на матрице ДНК) и трансляцию (синтез белков на матрице РНК). Если в том же участке хромосомы будет расположена другая последовательность триплетов, то будет производиться другой белок.

В генах хранится лишь информация о белках. Между работой генов и фенотипическими признаками лежит множество уровней биохимической организации. Производство некоторого белка может влиять на множество разных признаков (это так называемый *плейотропный* эффект). К примеру, некоторая мутация может приводить к синтезу белка, замедляющего скорость формирования волокон глиальных клеток, направляющих перемещение нейронов в процессе формирования мозга, что сложным образом скажется на его структуре. Лишь немногим генам может быть сопоставлен конкретный признак, такой как цвет горошин. В этом смысле Менделю повезло в том, что изучаемые им признаки были обусловлены отдельными генами, да еще и располагающимися в разных хромосомах. Ведь потом, когда он попытался воспроизвести свои опыты на других растениях, а также пчелах, результаты, полученные им на горохе, на других видах не подтверждались.

У разных видов последовательности нуклеотидных оснований могут быть совершенно разными. У особей одного вида хромосомы, как правило, имеют одно и то же разбиение на гены (т. е. можно сказать, что содержат одни и те же файлы, но с немного разным содержимым), при этом в каждом обособленном участке хромосомы (*локусе*) последовательность триплетов от особи к особи не меняется произвольным образом. Как правило, внутри вида в каждом локусе может быть сравнительно небольшое число разных вариантов последовательности. При этом геном принято называть сам выделенный участок хромосомы, а разные варианты последовательности триплетов (встречающиеся у особей данного вида) полагают разными формами этого гена, называемыми *аллелями*. У диплоидных организмов (в том числе человека) имеется двойной набор хромосом, поэтому даже у одной особи в генотипе может содержаться два разных варианта одного и того же гена (т. е. две аллели в соответствующем локусе пары гомологичных хромосом).

Генетическое разнообразие вида, вызванное тем, что локусы могут содержать разные аллели, сравнительно невелико. К примеру, для числа различных сочетаний генов человека иногда приводится цифра около 10^{50}. С одной стороны, это

число на много порядков превосходит число людей. Но, с другой стороны, для такого разнообразия достаточно всего лишь 166 генов, у каждого из которых есть две разные формы (иными словами, генетическое разнообразие людей описывается всего 166 битами или 21 байтом!). Возможно, эта оценка занижена, ведь у человека в геноме насчитывается порядка 25 тысяч генов, часть которых имеет разные формы (если бы каждый ген имел по две формы, то генетическое разнообразие описывалось бы 25 килобитами).

Объем информации, хранимый в геноме, конечно, гораздо больше. Гены могут включать тысячи и даже десятки тысяч нуклеотидных оснований. Всего в геноме человека около трех миллиардов (пар) оснований, или около миллиарда триплетов, каждый из которых кодирует одну из 20 аминокислот. В этом смысле полный объем генома можно оценить в 4 гигабита или 500 мегабайт (если эту оценку делать по числу пар оснований, то она будет несколько больше).

Насколько эта оценка завышена — сказать сложно. Ведь существуют гораздо более простые животные с гораздо более объемным геномом. Таким образом, геном (в том числе и человека) может быть информационно очень избыточным. Это как будто подтверждается тем, что участки, кодирующие какие-либо белки, составляют в геноме очень небольшой процент (кодирующие участки называются *экзонами*). Эволюционист Р. Докинз сравнивает некодирующие участки (называемые *интронами*) со стертыми файлами: само содержимое файлов осталось на диске, но в файловой системе соответствующие области помечены как неиспользуемые. Это сравнение подтверждается также и тем, что гены часто состоят из нескольких экзонов. Такое хранение генетической информации подозрительно напоминает фрагментацию диска: как будто файл не удалось записать на один свободный участок, и его пришлось разбивать на фрагменты и записывать в разные места. Если бы интроны соответствовали «стертым» генам, то оценку содержащейся в геноме информации можно было бы уменьшить на два порядка, однако в этой «мусорной ДНК» есть участки, выполняющие важные функции, к примеру контроль над экспрессией других генов, что позволяет одному гену в разных тканях

производить разные белки. Интроны — это «изобретение» многоклеточных организмов, необходимое для того, чтобы клетки с одним генотипом могли в организме выполнять разные функции (для них формула «один ген — один белок (или одна молекула РНК)» оказывается слишком упрощенной). Это подчеркивается и тем, что у прокариот интронов как будто нет. Однако даже если не вся «мусорная ДНК» является бесполезной, избыточность генома также видна из того факта, что отдельные гены в нем дублируются многократно в разных участках. Количественно избыточность сложно определить, но можно ли установить, какая часть генов связана с интеллектом?

Некоторые гены экспрессируются только в определенных органах и в определенных условиях. В частности, установлено, что у человека около половины генов экспрессируются *только* в нервной системе (так называемые нейрогены), а в других клетках — «молчат». Даже у крыс доля таких генов превышает 30 %. Часть этих генов активируется в эмбриогенезе при формировании структуры мозга, а часть — в течение жизни, особенно в процессе обучения и работы механизмов долговременной памяти. Можно полагать, что интеллект описывается значительной частью генов.

Стоит также отметить, что исходные данные для воссоздания человека не ограничиваются лишь тем конечным объемом дискретной информации, которая записана в ДНК. Ведь человек развивается из клетки, которая сама по себе является нетривиальной самовоспроизводящейся «машиной», задающей представление генетической информации. Остается лишь надеяться, что неопределенный объем информации, содержащейся в самой живой клетке, имеет очень косвенное отношение к интеллекту. Сейчас открываются все новые эпигенетические механизмы наследования, которые не затрагивают последовательности ДНК. В связи со всем этим истинную избыточность генома и объем наследственной информации об интеллекте оценить весьма сложно.

Итак, по пессимистичной оценке разработчикам ИИ придется писать сотни мегабайтов кода, чтобы их ИИ дальше смог уже самостоятельно развиваться. Это обозримый, но недоступный для маленьких исследовательских групп объем.

Данный объем, видимо, сильно завышен, но насколько именно — сложно сказать. По умеренно-оптимистичной оценке это значение нужно уменьшить порядка на два. Генетические исследования дают пока весьма приблизительные оценки объема наследственной информации об интеллекте. А какова при этом доля приобретенной информации?

Мозг в течение своей жизни обрабатывает заметно больший объем информации, чем содержится в геноме. Достаточно сказать, что в году 31 536 000 секунд, и каждую секунду в мозг поступают сотни мегабайт информации. Конечно, поступаемая информация обладает чрезвычайно большой избыточностью, да и просто может быть мало полезной для развития разума. Если же принимать в расчет только осмысленную информацию, то ее объем будет в грубом приближении сопоставим с генетической.

Для человека вопрос о степени наследуемости интеллекта имеет более конкретный смысл и большое прагматическое значение. Может ли у умных родителей родиться ребенок с низкими способностями? Имеет ли смысл пытаться развивать его интеллект? Конечно, все люди рождаются разными — по росту, весу, цвету волос и глаз и т. д. И было бы странно ожидать, чтобы мозг у них был идентичным. Естественно, люди давно замечали не только внешнее сходство между детьми и родителями, но также сходство их темперамента и прочих особенностей поведения.

Упоминавшийся уже Фрэнсис Гальтон (который был, кстати, двоюродным братом Дарвина) еще во второй половине XIX века, проанализировав множество родословных знаменитых людей, показал, что родственники выдающихся людей гораздо чаще также оказываются выдающимися, причем эта связь тем сильнее, чем теснее родство. Известны семьи, такие как Бахи, Бернулли, Ляпуновы, в роду которых были десятки талантливых людей. Часто эти факты рассматривают как свидетельство того, что талант — сугубо наследственное свойство. Конечно, исследователи вполне понимают, что дети, родившиеся в семье талантливых людей, растут в иных условиях, которые и могут быть причиной их успехов. В конце концов, почему сходство в музыкальных способностях между родителями и детьми считается

наследственным, а в религиозности или политических взглядах — нет?

Влияние как наследственности, так и среды, на способности очевидно. Однако часто делается следующая ошибка: считается, что это влияние аддитивно, т. е. врожденный интеллект как бы складывается с приобретенным. И если ребенок родился с «лучшими» генами, то он обгонит в развитии ребенка с более «плохими» генами, если этих детей поместить в одинаковые условия. Делается вывод, что улучшение (обогащение) среды может повысить интеллект обоих детей, но второй ребенок никогда не обгонит первого. Это заблуждение, верное лишь для утрированных случаев. В действительности все гораздо сложнее. Упрощенное понимание роли наследственности и среды может приводить к опасным педагогическим, социальным и политическим последствиям.

Гальтоном на основе своих исследований в 1865 году была сформулирована идея улучшения природы человека через размножение талантливых людей. Сходные идеи высказывались и другими мыслителями, например В. М. Флоринским в этом же году была опубликована статья «Усовершенствование и вырождение человеческого рода». Развитие этих идей привело Гальтона к основанию в 1883 году новой науки об улучшении человеческой породы — *евгеники* (буквально означающей «наука о рождении лучших»).

Некоторое время евгеника оставалась полезным научным направлением, в рамках которого копились сведения об особенностях наследования психологических признаков. Однако с начала XX века в разных странах стали активно внедряться евгенические программы. Наиболее «прогрессивными» оказались США, где уже в 1907 году в одном штате (а к 1937 году — в 32 штатах) был принят закон о принудительной стерилизации людей некоторых групп (к примеру, страдающих эпилепсией, хроническим алкоголизмом, умственной недостаточностью, психическими заболеваниями и т. д.). Более того, создавалась система запрета на иммиграцию в страну для национальностей, которые, предположительно, могут ухудшить наследственность американцев. Евгенические законы были приняты и в Германии.

Начав со спорных законов о запрете на потомство для людей с социально неблагоприятной наследственностью, нацистское руководство затем провозгласило уже откровенно ложные идеи национального превосходства, на основе которых «оправдало» массовые убийства.

В результате отношение к евгенике в целом стало предвзятым, и после Второй мировой войны обширные евгенические программы были закрыты. При этом в разных странах сохранились законы, которые вполне подходят под отдельные пункты евгенических программ начала XX века: управление размером семьи (как стимулирующие, так и ограничивающие меры — в зависимости от страны), поощрение или препятствие иммиграции в зависимости от социального статуса, профессии, наличия научной степени и т. д., стерилизация (часто — на выбор взамен другого наказания). Эти законы, однако, перестали быть объединены идеей улучшения генофонда.

Сама эта идея является не столь плохой, ведь в условиях отсутствия естественного отбора, усугубляющегося современной медициной, происходит накопление генетических дефектов. Однако в момент выдвижения данной идеи не было этически приемлемых средств для ее осуществления. Да и не было достаточно знаний о том, когда именно следует применять соответствующие меры. Полного решения этих проблем нет до сих пор. К примеру, известно, что алкоголизм или шизофрения в значительной степени обусловлены генетически, но люди с неблагоприятными генами вполне могут избежать этих заболеваний. Еще сложнее вопрос с уровнем наследования криминальных склонностей. Насколько вправе общество лишать подобных людей возможности иметь детей? Можно ли быть уверенным, что соответствующие гены (если таковые существуют) в благоприятных условиях не дадут обществу гениальных поэтов, художников или ученых? Современная генная инженерия лишь приближается к созданию гуманных средств устранения явных генетических дефектов, а генетика поведения — к ответу на поставленные вопросы.

Сейчас специалисты по психогенетике придают всем этим рассуждениям более строгий, количественный смысл, про-

водя исследования, в которых бы разделялись средовые и генетические факторы. Для этих целей может использоваться сравнение уровня интеллекта (в форме IQ-тестов, академической успеваемости и т. д.) монозиготных близнецов (обладающих идентичными генотипами), выросших вместе (т. е. примерно в одинаковой среде) и порознь. Также анализируются данные по дизиготным близнецам и просто сибсам (братьям и сестрам). Сравнение обычно производится с использованием коэффициента корреляции, вычисляемого по формуле

$$C = \frac{\sum_i (I_{1,i} - I_{1,\text{ср}})(I_{2,i} - I_{2,\text{ср}})}{\sqrt{\sum_i (I_{1,i} - I_{1,\text{ср}})^2}\sqrt{\sum_i (I_{2,i} - I_{2,\text{ср}})^2}}.$$

Здесь, например, $I_{1,i}$ может быть уровень интеллекта родителя, а $I_{2,i}$ — его ребенка (i — номер семьи в выборке); $I_{1,\text{ср}}$ — средний уровень интеллекта родителей, а $I_{2,\text{ср}}$ — их детей, тогда C — коэффициент корреляции уровней интеллекта родителей и их детей.

Можно убедиться, что если $I_{1,i} = I_{2,i}$ для всех i, то значение коэффициента корреляции будет равно единице. Если же значения в парах никак не связаны, то коэффициент корреляции будет близок к нулю. Он также может быть отрицательным, если бóльшие значения $I_{1,i}$ будут соответствовать меньшим значениям $I_{2,i}$, и наоборот.

Оказывается, если посчитать значение коэффициента корреляции уровня интеллекта у монозиготных близнецов, выросших вместе, то оно получится около 0,85, а у выросших порознь — чуть больше 0,7. Для дизиготных близнецов, выросших вместе, это значение оказывается около 0,6. Для сибсов (имеющих в среднем 50 % общих генов), выросших вместе, коэффициент корреляции оказывается около 0,5 (интересно, что эта величина примерно соответствует также и уровню сходства по физическим признакам типа роста). Однако для братьев и сестер, выросших порознь (например, усыновленных в разные семьи), в разных исследованиях данные приводятся разные. В некоторых исследованиях обнаружено, что коэффициент корреляции для них также

оказывается около 0,5, тогда как в других исследованиях получена цифра около 0,25. Дети, выросшие вместе и не являющиеся родственниками, обладают коэффициентом корреляции уровня интеллекта чуть больше 0,3.

Как отсюда получить степень наследуемости интеллекта? Для этого можно, в частности, сравнить значения коэффициентов корреляции для монозиготных и дизиготных близнецов, выросших вместе. У первых 100 % общих генов, тогда как у вторых — 50 %, поэтому разница 0,85 − 0,6 = 0,25 обусловлена тем, что число общих генов в два раза изменяется. Значит, уровень наследуемости интеллекта составляет примерно $0{,}25 \cdot 2 = 0{,}5$, т. е. 50 % интеллекта обусловлено генами, а 50% — средой. Это достаточно типичная оценка. В разных исследованиях разными учеными получались разные значения, которые обычно попадают в диапазон 0,3–0,8. Различия эти связаны как с использованием разных методик, так и с тем, что уровень наследуемости зависит от возраста и других факторов.

Делать выводы на основе этой цифры, однако, опасно. Ведь само измерение уровня интеллекта является сомнительной процедурой. К примеру, скорость прохождения тестов в большей степени, а результативность — в меньшей степени связаны с генотипом; и это различие возрастает по мере усложнения тестов. Нужно также понимать, что корреляция показывает лишь то, в какой степени наследственность отвечает за диапазон (разброс) уровней интеллекта в среднем по выборке. Это значит, что ее нельзя применять к каждому индивидууму по отдельности.

Да и сам коэффициент корреляции — вещь коварная. К примеру, если мы попробуем оценить, в какой степени количество ног у человека определяется наследственностью, то получим следующее. Подавляющее большинство людей рождается с двумя ногами. Вариативность количества ног от генотипа человека стремится к нулю. В то же время из-за условий среды человек может потерять ногу. Средовая дисперсия количества ног у человека невелика, но и не нулевая. Таким образом, получаем, что вариативность числа ног у человека практически полностью определяется условиями среды, хотя ни у кого не вызывает сомнения, что наличие

двух ног у человека «записано» в генах. Почему получен такой результат? Естественно, это связано с тем, что рассматривался только человек. Если оценить вариативность числа ног у всех животных, то получим преимущественно генетическую обусловленность.

Посмотрим также на следующий пример. Коэффициент корреляции интеллекта детей с интеллектом родных матерей во многих исследованиях оценивается примерно как 0,45, а с приемными матерями — около 0,2. Значение коэффициента корреляции уменьшается, если вычисляется для детей, разлученных с биологическими матерями, но все же остается выше (иногда называется значение 0,25, а иногда и больше), чем с приемными матерями. Значит ли это, что приемная семья не может создать для ребенка среду, влияние которой перевесило бы влияние наследственности? В действительности, такой вывод сделать нельзя, ведь коэффициент корреляции считается через разброс относительно *среднего значения*. У приемной матери ребенка интеллект нередко выше, чем у его биологической матери. Но если у приемных детей интеллект поднимается в среднем, это не отражается на коэффициенте корреляции. Давайте взглянем на следующие цифры (условно здесь $I_{1,i}$ — IQ приемных детей; $I_{2,i}$ — IQ их биологических матерей; $I_{3,i}$ — IQ их приемных матерей):

$I_{1,i}$	112	107	115	110	101
$I_{2,i}$	84	92	93	88	83
$I_{3,i}$	107	122	116	112	108

На кого больше похожи в этой гипотетической ситуации приемные дети? Сами значения $I_{1,i}$ и $I_{3,i}$ кажутся ближе друг к другу. Но попробуем посчитать значения коэффициента корреляции. Для этого найдем средние значения: $I_{1,\text{ср}} = 109$; $I_{2,\text{ср}} = 88$; $I_{3,\text{ср}} = 113$ и посмотрим на отклонения от средних:

$I_{1,i}-I_{1,\text{ср}}$	+3	−2	+6	+1	−8
$I_{2,i}-I_{2,\text{ср}}$	−4	+4	+5	0	−5
$I_{3,i}-I_{3,\text{ср}}$	−6	+9	+3	−1	−5

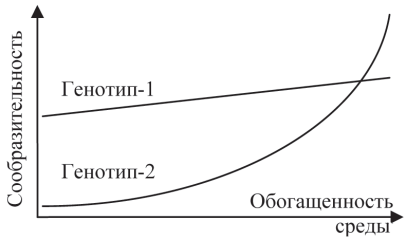
Вычислим значения коэффициентов корреляции: $C_{1,2} = 0{,}52$ и $C_{1,3} = 0{,}16$. Иными словами, средние значения IQ больше

похожи у детей с приемными матерями, а отклонения от среднего — с биологическими матерями. Хотя приведенные данные являются модельными, этот пример показывает, что такая ситуация в принципе может возникнуть.

Зачем же вычитается среднее? Дело в том, что изменение среднего значения может быть вызвано разными факторами. К примеру, значение IQ зависит от возраста, поэтому сравнивать средние значения для детей и взрослых в данном контексте не вполне корректно. Кроме того, повышение уровня интеллекта детей может быть вызвано общими факторами — доступностью образования, распространением компьютеров и т. д.

В случае исследования уровня наследования интеллекта человека все эти факторы плохо контролируются. Эксперименты с животными позволяют несколько прояснить ситуацию. В частности, в них можно выводить чистые линии (т. е. семейства, например крыс, обладающих идентичными генотипами) путем так называемого *имбридинга* — скрещивания близких родственников в рамках одной популяции. Когда генотип — константа, все разнообразие поведения будет определяться различиями условий среды. Для данного генотипа можно построить график зависимости некоторого признака (будь то густота шерсти или уровень интеллекта) от условий среды. Для этого животные чистой линии помещаются в разные условия, далее определяется доля животных, у которых данный признак принял некоторое значение. Оказывается, каждый генотип обладает индивидуальной зависимостью значения фенотипического признака от параметров среды, получившей название *нормы реакции*. Сейчас принято говорить, что наследуется не какой-то признак, а норма реакции.

Рассмотрим следующую типичную ситуацию. Выведены две чистые линии крыс, причем особи с одним генотипом отличаются в среднем большей «сообразительностью» (выражаемой, например, в скорости прохождения лабиринта), чем особи с другим генотипом. Однако часто оказывается так, что более «глупые» животные (Генотип-2), помещенные в обогащенную среду, умнеют, тогда как исходно более сообразительные животные (Генотип-1) остаются примерно на



Примеры разных норм реакций

том же уровне. Это можно представить следующими нормами реакции (см. рисунок).

Какой из этих двух генотипов более желательный? Один в среднем (для обычных сред) дает более умных животных, но другой обладает бóльшим, хотя и труднее реализуемым, потенциалом. Понятие нормы реакции показывает всю трудность проблем евгеники. Но и это понятие является сильно упрощенным. Проблема в том, что норма реакции не является абсолютной характеристикой — она зависит от рассматриваемого диапазона сред. Обычно норма реакции укладывается в некоторый ограниченный диапазон значений признака. Может сложиться впечатление, что за этот диапазон при данном генотипе невозможно выйти. Однако нельзя с уверенностью утверждать, что не существует такой среды с необычными свойствами, при взаимодействии с которой генотип вышел бы за свои стандартные рамки. В частности, изменение нормы реакций при повышении разнообразия пренатальных условий развития очень мало изучено.

Можно провести следующую аналогию (хотя, возможно, не вполне корректную). Чем менее развит интеллект у некоторого вида животных, тем менее беспомощными они рождаются. Многие низшие животные после рождения сразу готовы бегать, летать или даже инстинктивно использовать некий язык. Детеныши шимпанзе, растущие вместе с человеческими детьми, до двух лет обгоняют их в развитии. Нельзя гарантировать, что во всех случаях замедленное развитие ребенка означает более низкий интеллект и что быстрое раннее развитие приводит всегда к хорошему результату в будущем (процент вундеркиндов, достигших выдающихся результатов в профессиональной деятельности, гораздо мень-

ше, чем того следовало бы ожидать). В ряде случаев (при отсутствии патологий) замедленное развитие может означать, напротив, более универсальный интеллект. Такому интеллекту для становления может требоваться специальная среда и больший объем информации, но он получит контроль над этой информацией, способность ее критического переосмысления.

Как видно, взаимодействие генотипа и среды весьма сложное. Причем это взаимодействие затрагивает не только онтогенетическое развитие организма, но также и эпигенетические факторы наследования, действующие наряду с генетическим наследованием.

Исследования в психогенетике пока еще не дали полного ответа на вопрос о роли генотипа и обучения в формировании интеллекта, но уже послужили стимулом для развития таких областей, как эволюционная и эпигенетическая робототехника. Становится понятно, что проблемы эволюции и адаптивного поведения необходимо рассматривать совместно.

Здесь следует вспомнить биогенетический закон Э. Геккеля, в соответствии с которым онтогенез есть рекапитуляция (быстрое и краткое повторение) филогенеза, т. е. при индивидуальном развитии зародыши животных принимают форму предшествующих видов, начиная с наиболее древних. Ст. Холл и Дж. Болдуин расширили эту концепцию и на психологическое развитие: ребенок в сокращенной форме проходит путь эволюционного и культурного развития человечества. Этому закону неплохо соответствуют даже программы школьного и институтского образования: изучение предметов начинается с наиболее древних достижений (хотя и в заметной мере переработанных). Если развитие ИИ будет следовать тому же принципу, то ему также придется в каком-то виде повторять путь культурного развития человечества. В любом случае, поскольку значительная часть интеллекта обусловлена средой, а для ИИ в качестве среды будет выступать человеческий социум, то ИИ в рамках этого сценария может в некотором смысле трактоваться как следующий этап эволюции интеллекта — после биологического и социального.

Сходство индивидуального и эволюционного развития, на самом деле, является гораздо более фундаментальным, что

хорошо видно в рамках исследований ИИ. Действительно, чем виртуальная имитация эволюционного возникновения некоторой формы поведения принципиально отличается от машинного обучения? По сути, такая имитация — это форма поиска в пространстве решений.

ЭВОЛЮЦИОННЫЕ ВЫЧИСЛЕНИЯ

Насколько осмысленной является идея использования законов эволюции не при имитации самой эволюции, а при решении интеллектуальных задач? На первый взгляд, идея может показаться интересной, но немного сомнительной: не происходит же у нас в голове эволюция, пусть даже виртуальная. Однако эта идея была высказана достаточно давно сразу несколькими авторами и оказалась относительно успешной. Сейчас область исследований, опирающихся на сходные идеи, называется *эволюционными вычислениями*. Наиболее ранним вкладом в нее в 1960-х годах стали *эволюционные стратегии*, предложенные Инго Рехенбергом с коллегами, и эволюционное программирование, предложенное Лоуренсом Фогелем, а также *генетические алгоритмы* Джона Генри Холланда (которые стали известны в 1970-х). Позднее возникла концепция *генетического программирования*. Все эти разновидности эволюционных вычислений являются альтернативой классическим методам поиска и оптимизации, включая эвристическое программирование.

Существующие методы эволюционных вычислений берут за основу идеи Дарвина. Одна из этих идей гласит: выживает наиболее приспособленный. В действительности, это тавтология, поскольку более приспособленный — это, по определению, тот, кто имеет больше шансов выжить. Эту тавтологичность подчеркивают и сами эволюционисты, чтобы доказать неизбежность эволюции, несмотря на отсутствие у природы каких-либо специальных целей. Не случайно идея выживания наиболее приспособленных часто называется естественным отбором. Отбор обычно подразумевает целенаправленное действие, но в природе он происходит «автоматически».

Однако неизбежность эволюции не следует лишь из логической истинности естественного отбора. Ведь для того, чтобы эволюция стала «работать», необходимо появление новых альтернатив, из которых осуществляется отбор. В связи с этим вводится дополнительная идея — идея изменчивости видов. То, что виды изменяются со временем, не самоочевидно и может быть установлено только из эмпирических данных, поскольку изменчивость должна обеспечиваться какими-то конкретными механизмами. Так, мы не можем сказать, что сейчас на уровне элементарных частиц идет естественный отбор, поскольку не происходит возникновение все новых и новых видов частиц.

Еще одна идея — это идея наследственности. Если бы виды возникали произвольно, то со временем могли бы появляться все более приспособленные виды, однако появление новых видов не зависело бы от степени приспособленности уже существующих, ведь при возникновении нового вида заранее нельзя сказать, насколько приспособленным он будет (если этот вид не конструируется сознательно). Наследственность же обеспечивает последовательное улучшение существующих решений за счет изменчивости и естественного отбора (стоит отметить, что здесь неявно присутствует предположение непрерывности: небольшое изменение вида обычно не сильно сказывается на степени его приспособленности).

Таким образом, как хорошо известно, базовые элементы дарвиновской концепции эволюции — это наследственность, изменчивость и естественный отбор. Поскольку в процессе эволюции появляются все более приспособленные виды, эволюцию можно трактовать как поиск максимума некоторой функции выживания, или *фитнесс-функции*. Если предположить, что в процессе эволюции в некоторый момент создан совершенный вид, то на нем эволюция прекращается, так как изменение этого вида может привести только к ухудшению его выживаемости (т. е. к уменьшению значения фитнесс-функции). Оставим на время в стороне вопрос о том, насколько биологическая эволюция соответствует такой модели, и согласимся с определенным сходством между эволюцией и поиском, которое и было замечено исследователями в области ИИ.

Однако перечисленные выше идеи Дарвина добавляют немногое к классическим методам поиска. Чтобы показать это, достаточно переформулировать упоминавшийся при обсуждении проблем поиска метод градиентного спуска (или подъема в гору, т. е. движения в направлении наискорейшего локального возрастания или убывания функции) в эволюционных терминах. Пусть текущая точка в методе градиентного спуска — это некоторая особь. На каждом шаге градиентного спуска проверяются некоторые точки вблизи текущей (наследственность и изменчивость) и выбирается из них лучшая (отбор). Стохастический градиентный спуск из многих точек обладает еще большим сходством с дарвиновской эволюцией, если ее ограничить лишь указанными идеями. На самом деле, еще сам Дарвин называл отбор в сочетании с изменчивостью «спуском с модификацией». Таким образом, обычный градиентный спуск содержит все базовые элементы эволюции. С одной стороны, это еще раз подчеркивает сходство эволюции с поиском, но с другой стороны, ставит вопрос, что же исследователи ИИ нашли для себя нового в эволюции?

Заинтересовали их не только базовые идеи Дарвина, но и генетические механизмы передачи наследственной информации. В качестве таких механизмов чаще всего рассматриваются перекрест хромосом при скрещивании и генные мутации.

В общем виде алгоритмы эволюционных вычислений, предназначенные для поиска оптимального решения некоторой задачи, могут быть представлены следующим образом. Альтернативные решения (или оптимизируемые объекты) трактуются как особи, степень приспособленности которых, или фитнесс-функция, явно или неявно определяется условиями задачи. Эти особи «эволюционируют» — к ним применяются *«генетические операторы»*, такие как операторы *скрещивания, мутации* и *редукции* (селекции или отбора). Такие алгоритмы обычно состоят из следующих шагов.

1. Сгенерировать начальную популяцию (случайную совокупность неоптимальных решений).

2. Выбрать родительские пары.

3. Для каждой родительской пары с использованием оператора скрещивания породить потомство.

4. К порожденным особям применить оператор мутации, внеся случайные искажения.

5. Произвести отбор особей из популяции по значению их фитнесс-функции, применив оператор редукции.

6. Повторять шаги 2–5, пока не выполнится критерий остановки.

Каждый шаг имеет разные реализации. Кроме того, особенности генетических операторов зависят от конкретной формы эволюционных вычислений. Так, особенностью генетических алгоритмов является то, что в них для представления решений используются битовые строки, трактуемые как генотипы (обычно состоящие из единственной хромосомы), и все генетические операторы применяются к битовым строкам. В эволюционных стратегиях операторы применяются к самим решениям, представленным в их естественной форме. Особенность же эволюционного (генетического) программирования состоит в том, что в них рассматривается оптимизация особых объектов — компьютерных программ. Рассмотрим каждый из шагов чуть подробнее на примере генетических алгоритмов (ГА).

1. Генерация начальной популяции обычно производится равномерно по пространству генотипов. Размер популяции — установочный параметр.

2. Выбор родительских пар может осуществляться различными способами. Обычно он включает два этапа: выбор первого родителя и формирование пары. При выборе первого родителя обычно используется один из следующих способов:

- с равной вероятностью выбирается любая особь из имеющейся популяции;
- особь выбирается случайно с вероятностью, пропорциональной значению фитнесс-функции; т. е. в этом случае значение фитнесс-функции сказывается не только на том, какие особи останутся в популяции в результате отбора, но и на том, сколько потомства они произведут.

Выбор второго родителя осуществляется по одному из следующих критериев:

• независимо от уже выбранного родителя (т. е. второй родитель выбирается абсолютно так же, как и первый); этот вид отбора называется *неселективным*;

• на основе *ближнего родства*;

• на основе *дальнего родства*.

В последних двух случаях выбор одного родителя влияет на выбор другого родителя: с большей вероятностью формируются пары, состоящие из особей, которые больше похожи друг на друга (т. е. ближе находятся в пространстве генотипов) при использовании ближнего родства или меньше похожи при использовании дальнего родства. В генетических алгоритмах в качестве меры близости обычно используется расстояние Хемминга, которое для двух битовых строк вычисляется как число позиций, в которых в двух строках стоят несовпадающие символы (т. е. в одной строке стоит 0, а в другой — 1).

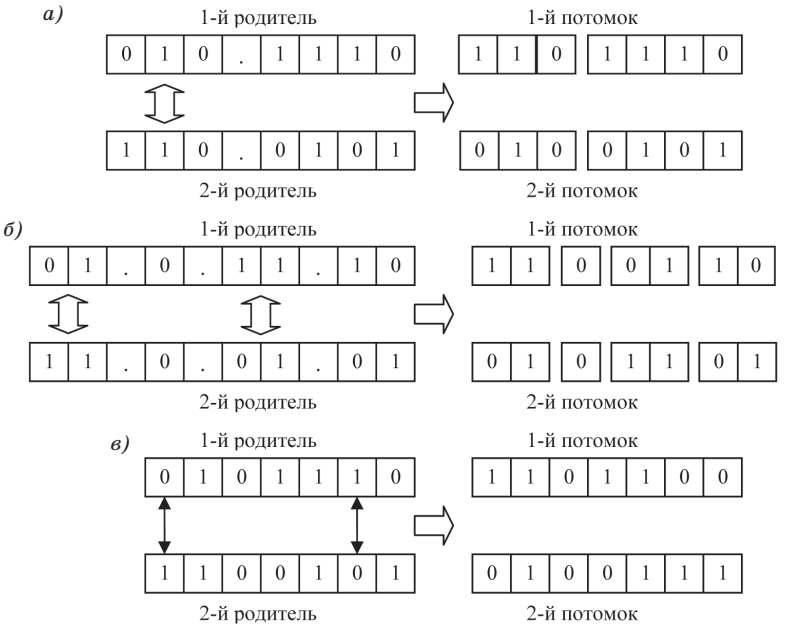
3. Оператор скрещивания — это оператор, который определяет, как из генотипов родителей формировать генотипы их потомства. Один из интуитивно очевидных способов заключается в том, чтобы каждый бит генотипа для потомка брать от случайного родителя. Такой способ действительно используется и называется *равномерным скрещиванием*. Однако в природе гены внутри одной хромосомы являются *сцепленными*, поэтому случайным (и независимым) образом от родителей берутся целые хромосомы. Также есть механизм, который позволяет рекомбинировать и сцепленным генам. Это механизм *кроссинговера*, или перекреста хромосом, при котором (гомологичные) хромосомы обмениваются участками.

В ГА этот процесс моделируется так: хромосомы (а как отмечалось выше, обычно в ГА все гены особи располагаются в единственной хромосоме) делятся в некоторой случайной точке и обмениваются этими участками (т. е. все, что идет до этой точки, берется от одного родителя, а все, что после, — от другого). Это одноточечный кроссинговер. В многоточечном кроссинговере таких участков обмена больше. В ГА нередко используется такой оператор скрещивания, при котором формируются генотипы сразу двух потомков, содержащие всю генетическую информацию родителей. Варианты реализации оператора скрещивания представлены ниже на рисунке.

Если бы в ГА геном представлялся в виде нескольких хромосом, то моделировался бы одновременно случайный выбор целых хромосом и рекомбинация генов внутри отдельных хромосом с помощью кроссинговера.

4. В бытовом понимании мутации обычно представляются как случайные искажения генов, и именно так (несмотря на то, что это крайне упрощенное представление) они реализуются в ГА. Действие оператора мутации сводится к случайной замене одного (иногда нескольких) бита генотипа. Настроечный параметр алгоритма — скорость мутации — определяет, как часто эта замена делается. Этот параметр влияет на скорость сходимости и вероятность попадания в локальный экстремум. Естественно, чем чаще мутации, тем медленнее стабилизируется генофонд популяции, но тем меньше шансов, что эволюция «застрянет» в неоптимальном решении.

5. Отбор особей в новую популяцию чаще всего осуществляется в соответствии с одной из двух стратегий:



Примеры работы оператора скрещивания: *а* — одноточечного; *б* — многоточечного; *в* — обмена случайными битами

• пропорционального отбора, при котором вероятность того, что особь останется в следующей популяции, пропорциональна значению ее фитнесс-функции;

• элитного отбора, при котором из популяции отбираются лучшие по значению фитнесс-функции особи, и только они переходят в следующую популяцию.

Формирование новой популяции может осуществляться как на основе потомков и родителей, так и на основе только потомков в зависимости от конкретной реализации.

6. Основные критерии остановки алгоритма базируются либо на числе сменившихся поколений (количестве выполненных итераций), либо на некотором условии стабильности популяции. Заранее сложно предсказать, сколько именно популяций потребуется для сходимости, поэтому этот критерий используется обычно как вспомогательный. Проверка стабильности популяции в общем виде, как правило, требует значительных вычислений, поэтому чаще используется проверка того, что наилучшее по популяции значение фитнесс-функции перестает заметно изменяться от поколения к поколению.

Отличие эволюционных стратегий от генетических алгоритмов заключается в том, что в первых не используются битовые представления. Вместо этого все генетические операторы реализуются в пространстве исходных объектов (или фенотипических признаков) с учетом их структуры. Рассмотрим особенности реализации генетических операторов в эволюционных стратегиях на примере объектов, описаниями которых являются двухкомпонентные векторы вида (x, y), т. е. задача заключается просто в поиске экстремума функции от двух переменных $f(x, y)$.

1. Генерация начальной популяции может осуществляться путем выбора случайных векторов из прямоугольной области $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$, в которой ожидается нахождение экстремума фитнесс-функции. В случае генетических алгоритмов эта область задается неявно, и она зависит от способа отображения вектора (x, y) в битовую строку.

2. При выборе родителей особенность эволюционных стратегий выражается в способе задания меры родства. В данном случае мерой родства двух особей (x_1, y_1) и (x_2, y_2) может

служить евклидово расстояние, которое будет заметно отличаться от расстояния Хемминга, используемого в генетических алгоритмах.

3. Результатом скрещивания двух особей в рассматриваемом случае будет являться особь, находящаяся в случайном месте отрезка (x_1, y_1)—(x_2, y_2), что, опять же, отличается от результата скрещивания в пространстве генотипов.

4. Результатом мутации особи (x, y) будет являться особь $(x + \delta_x, y + \delta_y)$, где δ_x, δ_y — случайные величины, разброс которых определяет скорость мутаций.

5, 6. Операторы отбора и критерии останова в эволюционных стратегиях не имеют особых отличий от тех, которые используются в генетических алгоритмах.

Как видно, реализация генетических операторов на уровне фенотипов допускает более гибкую их настройку, что может оказать помощь в повышении эффективности поиска, но при этом требуются дополнительные усилия со стороны разработчика. При использовании ГА реализация операторов может быть одинаковой для разных задач, для которых нужны только процедуры перевода объектов в битовые строки и обратно. Однако неудачный способ перевода может привести к низкой эффективности ГА.

Оба этих подхода могут применяться при решении самых разнообразных задач — от поиска экстремума функций вещественных переменных до решения изобретательских задач по разработке конструкций технических систем (известны случаи даже патентования и коммерческого использования решений, найденных методами эволюционных вычислений). Они позволяют приближенно решать некоторые NP-полные задачи не хуже, чем методы эвристического программирования, и требуют при этом меньших усилий от разработчика.

Если оба подхода одинаково применимы на практике, почему в природе используется только один из них? Действительно, эволюционные стратегии в чем-то ближе к концепции не Дарвина, а Ламарка, согласно которой совершенствуются и непосредственно наследуются фенотипические признаки. Эта концепция неплохо работает, когда имеется фиксированный набор фенотипических признаков, для ко-

торых можно определить, как их изменения сказываются на фитнесс-функции. Но если оптимизируемые объекты комбинаторно конструируются, перечень их внешних признаков может быть неограниченным и заранее неизвестным. В этом случае применить поиск в пространстве фенотипов не представляется возможным. Именно такая ситуация имеет место в живой природе. Выше мы затрагивали вопрос о сложной связи генов и фенотипических признаков, которые разделяют многие уровни биохимического конструирования, способные породить неограниченный набор внешних признаков. Но за это приходится «платить» тем, что обратный переход от фенотипа к генотипу становится, по крайней мере, NP-полной задачей высокой размерности. Как результат, наследование приобретенных признаков в стиле Ламарка оказывается почти (но не полностью) невозможным.

При использовании генетических алгоритмов разработчик, однако, обычно начинает с фиксированных фенотипических признаков, данных в рамках решаемой оптимизационной задачи, и для них выдумывает способ кодирования в геноме одновременно с обратным преобразованием. Поскольку и в ГА фенотипы оказываются первичными, эволюционные стратегии имеют не меньше прав на применение.

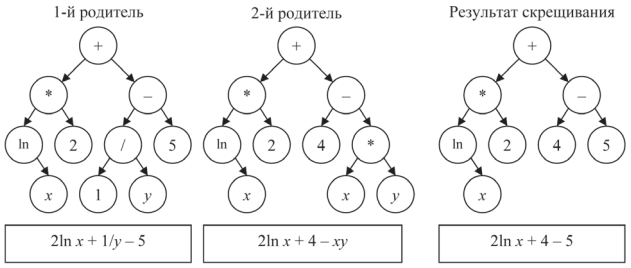
Существует, однако, класс объектов, эволюционная оптимизация которых затруднительна в пространстве их «фенотипических признаков», — это алгоритмы, или программы. Не существует какого-нибудь простого отображения между желаемым выводом программы и ее кодом (выполнение такого отображения — это NP-полная или алгоритмически неразрешимая, в зависимости от деталей постановки, задача). Не удивительно, что попытки автоматического построения программ в рамках эволюционных вычислений выделяются в самостоятельное направление — генетическое (эволюционное) программирование.

Здесь могут применяться аналоги как генетических алгоритмов, так и эволюционных стратегий. В обоих случаях перебираются программы, а не совершаемые ими действия, поэтому различия между ними сводятся к тому, в каком виде представлять код программы — в виде бинарной строки или, скажем, в виде графа. Из-за этого и трудности у этих

методов общие, хотя и выражающиеся немного по-разному, например, как реализовать операции скрещивания и мутации, чтобы после их выполнения получались корректные программы (или как закодировать программы в форме бинарной строки для достижения той же цели).

Чаще используются не программы на каком-то обычном языке программирования, а разрабатывается некий язык со специальным синтаксисом, упрощающий работу в рамках эволюционных методов. На практике этот язык зачастую не является алгоритмически полным. В простейшем случае (исходно рассмотренном Фогелем) структура программы фиксирована, а оптимизируются только ее параметры. В современных методах эволюционного программирования рассматриваются представления программ в виде деревьев. При этом в результате скрещивания одно поддерево заменяется другим (все родительские узлы у них должны совпадать). Пример простой (с точки зрения представления программы), но часто встречающейся задачи, — это подбор математического выражения под известные результаты вычислений. На рисунке ниже представлен один из возможных результатов скрещивания двух программ-выражений.

Такой способ поиска в пространстве программ хорошо сочетается с подходом к индуктивному выводу на основе алгоритмической сложности. Действительно, если у нас есть некоторый набор данных, для которого нужно построить модель, мы вполне можем применить эволюционное (или генетическое) программирование, при котором фитнесс-функция каждой особи-программы будет оцениваться по критерию минимальной длины описания. Не даст ли это универсального решения задачи индуктивного вывода? Та-



Скрещивание выражений

кая возможность, в числе прочих, исследовалась Р. Соломоновым. К сожалению, существующие методы эволюционных вычислений не справляются с NP-полнотой данной проблемы и позволяют получить приемлемое решение лишь для достаточно простых случаев.

Пожалуй, в приложении к проблемам машинного обучения наиболее популярным оказалось применение эволюционного программирования при обучении нейронных сетей, что в большей степени используется не в индуктивном выводе, а при синтезе систем управления, в том числе в «Искусственной жизни» и «Адаптивном поведении». Наиболее простым в реализации является случай, когда структура нейронной сети фиксирована и требуется лишь настроить веса ее связей. В геноме тогда кодируются только веса связей, так что длина генотипов оказывается одинаковой у всех особей и не возникает опасности порождения некорректного генетического кода в результате применения генетических операторов. Это, однако, приводит к сильному ограничению пространства поиска. Поиск на менее ограниченном множестве программ является все еще плохо изученной проблемой.

Одна из причин, почему разработчики пытаются ограничиться геномом фиксированного размера, заключается в том, что в противном случае не вполне ясно, как реализовывать оператор скрещивания. Стоит отметить, что фиксированная структура генома приводит к тому, что в эволюционных вычислениях обычно скрещиваются все особи, т. е. эволюционирует один вид. Возможность изменения структуры генома в ходе эволюционных вычислений с соответствующим расщеплением видов практически не исследована.

Порождение ИНС методами эволюционных вычислений лишь заменяет традиционные методы их обучения. В этой связи гораздо больший интерес может представлять порождение нейроглиальных сетей, ведь в применении к ИНГС эволюционные вычисления превращаются в метаобучение — автоматическое построение самих алгоритмов обучения. Пока, однако, в этом направлении сделаны лишь первые шаги.

Альтернативой эволюционного синтеза нейросетевых систем управления является порождение управляющих конечных автоматов (как классических, так и на основе теории

нечетких множеств), где также достигнуты определенные успехи.

Итак, эволюционные вычисления неплохо себя зарекомендовали в качестве методов поиска как при решении задач классического ИИ (т. е. при решении корректно поставленных NP-полных задач), так и при решении задач машинного обучения, хотя и не дали их полного решения.

Почему же методы эволюционных вычислений оказываются достаточно успешными для решения проблем поиска? Эволюционные стратегии в применении к поиску экстремума функции от вещественных переменных очень похожи на стохастический градиентный спуск за исключением того, что в них происходит «скрещивание» решений. С генетическими алгоритмами ситуация чуть сложнее, так как в них небольшие мутации генов могут привести к значительному изменению «фенотипа» (примером может служить изменение старшего бита в двоичной записи числа). Но и в ГА основной особенностью является скрещивание, благодаря которому новые решения строятся как фрагменты имеющихся решений. Если задача разбивается на подзадачи и решению каждой подзадачи соответствует отдельный участок генома, то за счет скрещивания подзадачи могут решаться независимо. Действительно, если решение какой-то подзадачи улучшает общее решение независимо от других подзадач, то в генофонде популяции соответствующий участок генома быстро стабилизируется (его вариативность сильно уменьшится). Еще в большей степени этому будет способствовать использование не равномерного скрещивания, а кроссинговера, при котором потомкам передаются длинные фрагменты генотипов родителей. Кроссинговер будет работать хорошо, только если решения отдельных подзадач локализованы в геноме.

Итак, мы видим, что ГА дополнены одним, но очень мощным приемом, благодаря которому решаемая задача «мягко» (но не адаптивно) разбивается на подзадачи. Следует, однако, иметь в виду, что этот прием не вполне универсален. Если участки генома не соответствуют фрагментам решений почти независимых подзадач (в силу особенностей решаемой NP-полной задачи или в силу неудачного выбора способа кодирования решения в геноме), то ГА будут смесью полного пере-

бора и градиентного спуска, причем реализованного не самым эффективным образом. Отдельные участки генома не будут стабилизироваться независимо друг от друга, и найти решение удастся, только если все гены одновременно рекомбинируют нужным образом, вероятность чего будет очень низка.

Прием, который позволяет эффективно сокращать перебор для многих задач, но в общем случае не гарантирует нахождения оптимального решения, принято называть *метаэвристикой*. В отличие от простой эвристики метаэвристика не является предметно-специфичной, что делает ее широко применимой, но во многих случаях недостаточной. К примеру, сами по себе ГА не могут использоваться при выборе хода в шахматах, для которых классическое эвристическое программирование пока остается хоть и весьма трудоемким, но практически единственным подходящим средством. Если же в какой-то NP-полной задаче части решения являются почти независимыми (как, например, в задаче коммивояжера), то единственная метаэвристика, заложенная в ГА, может оказаться достаточной для получения приемлемого решения, благодаря чему не нужно будет разрабатывать индивидуальный (под данную задачу) алгоритм обхода дерева варианта со сложными предметно-специфичными эвристиками.

Часто говорят, что эволюционные вычисления весьма эффективны, поскольку они заимствуют основные идеи из естественной эволюции. Но зададимся наивным вопросом: почему генетическое программирование не позволяет создать ИИ? Выше мы не напрасно останавливались на вопросах устройства генов: теперь мы видим, что ГА и аналогичные им методы столь же проще реальных генетических механизмов, сколь и ИНС проще биологических нейронов. Вместе с тем ГА, как и методы обучения ИНС, недостаточны для работы в алгоритмически полном пространстве.

Несмотря на некоторые положительные свойства, методы эволюционных вычислений оказываются вовсе не универсальными. В связи с этим, если мы рассмотрим эволюционные вычисления как методы поиска в пространстве решений, то становится понятным, почему с их помощью не так просто автоматически получить ИИ. По сути, эти методы сами будут составлять интеллектуальную систему, перед которой ставится задача поиска программы ИИ. Нет ли здесь противоречия? Не получается ли так, что нам уже нужно иметь ИИ, чтобы компьютер смог сам его придумать? Принципиального противоречия здесь нет: более «глупая» программа может вывести более «умную» посредством «грубой силы», т. е. используя очень интенсивный перебор. Конечно, чем глупее эволюционная программа, тем больше грубой силы ей придется приложить. Полным перебором задачу построения ИИ или даже игры в шахматы на практике не решить. Естественно, недостаточно и нескольких простых метаэвристик. Такой взгляд показывает наивность попыток с помощью простой искусственной эволюции вывести интеллектуальные программы. Но как это удалось сделать естественной эволюции?

ЭВОЛЮЦИЯ КАК ПОИСК

Дети в определенный момент начинают задаваться вопросом, откуда они появились, и через некоторое время получают на него ответ. Однако после этого интуитивные индуктивные рассуждения приводят к следующему вопросу: откуда появились люди вообще? — на который дать ответ уже гораздо сложнее. То, как просто к этому вопросу прийти и как сложно на него ответить, сделало из него предмет споров, не прекращающихся на протяжении многих веков. Свои ответы на него, в первую очередь, дают разные религии. Эти ответы, несмотря на некоторые различия, можно свести к идее креационизма, согласно которой человек (да и весь мир) создан неким Творцом, или Богом. Здесь нас, однако, интересует не столько сам человек, сколько его разум. На вопрос, как возник разум, креационизм отвечает, что он был создан еще более мощным разумом. Оставляя в стороне достоверность этого ответа вместе с вопросом, откуда взялся сам более мощный разум, мы лишь отметим, что такой ответ мало полезен исследователям ИИ. Гораздо больший интерес может представлять ответ, даваемый эволюционной теорией.

А мог ли разум действительно возникнуть в процессе эволюции? Сомнение в такой возможности высказывается нередко. Контраргумент, особенно любимый креационистами, заключается в следующем. В геноме человека три миллиарда пар нуклеотидных оснований. Чтобы случайным образом возник такой геном, необходимо, чтобы кубик с четырьмя гранями при подбрасывании три миллиарда раз выпал нужной стороной. Вероятность этого события $4^{-3\,000\,000\,000}$ является исчезающе малой. Даже с учетом того, что геном человека обладает определенной вариативностью и избыточностью, вероятность появления человека, рассчитанная таким способом, останется невообразимо низкой. Конечно, то, что возник именно человек, можно считать случайностью. В частности, наиболее развитые птицы считаются сопоставимыми по уровню интеллекта с низшими обезьянами, имея при этом другую организацию мозга: новая кора у них развита слабо, а большее развитие получил стриатум. Но даже если попробовать посчитать вероятность случайного возникновения произвольного разума, цифра окажется нереалистично маленькой. Мы не знаем минимального размера алгоритма автономного интеллекта, но даже если предположить, что он умещается в каких-то десять килобайт (это, скорее всего, весьма заниженная оценка), то вероятность того, что один случайный набор бит даст нужную последовательность будет, $2^{-81\,920} \approx 10^{-25\,000}$. Даже если учесть гипотетическое количество живых существ, живших на всех планетах нашей Вселенной с момента ее возникновения, то вероятность в таком количестве вариантов генотипа встретить нужную для реализации разума последовательность будет почти нулевая. Креационисты делают вывод: разум случайно возникнуть не мог. С этим выводом практически не поспорить, хотя многие материалисты часто от него просто отмахиваются, не задумываясь о том, что действительно почти нулевая вероятность случайного возникновения разума сама собой реализоваться не может.

Есть, однако, одно возражение, называемое *антропным принципом*. Основная идея этого принципа заключается в том, что если бы разум не появился, то никто бы не смог задаваться вопросом о вероятности его возникновения. Любой

разум сможет увидеть вокруг себя только такую реальность, в которой он мог бы возникнуть. Иными словами, априорная вероятность возникновения разума (которая может быть сколь угодно малой) не имеет значения; факт наличия разума нам уже дан, и его апостериорная вероятность равна единице. Только эта вероятность имеет значение. С этим выводом тоже не поспорить.

Антропный принцип в своей наименьшей формулировке оставляет некоторую неудовлетворенность: он не отвечает на вопрос, как возник разум, а лишь говорит, что раз уж разум есть, то Вселенная должна быть таковой, чтобы он был. Антропный принцип настолько безразличен к способу возникновения разума, что ни в коей мере не отвергает и креационизм (а лишь говорит о том, что креационизм ничуть не лучше гипотезы случайного возникновения), а также не доказывает эволюцию (лишь спасает ее от аргумента низкой вероятности). Не удивительно, что сторонники случайного возникновения разума (вернее, противники креационизма) дополняют его гипотетическими механизмами. Например, они говорят, что вселенных может быть практически бесконечное число (или наша Вселенная прошла бесконечное число циклов от Большого взрыва до коллапса), и лишь в тех немногих из них, где возник разум, разумные существа имели сомнительное счастье удивляться факту своего маловероятного существования. По сути, такая гипотеза «многих попыток» к антропному принципу не имеет никакого отношения, поскольку она пытается объяснить факт существования разума за счет того, что разум имеет высокую вероятность возникновения *хотя бы в одном из миров*, если этих миров взять достаточно много.

Такое объяснение сходится с позицией креационистов в том, что вероятность чисто случайного возникновения разума ничтожно мала. И с этим, конечно же, можно согласиться. Однако это вовсе не значит, что у нас останется две альтернативы: либо выдумывать гипотетические возможности (типа бесчисленного множества пустых миров) чисто случайного возникновения разума, либо полагать, что он возник сверхъестественным путем. В действительности, одновременно и идее случайного возникновения разума, и

идее его сотворения можно противопоставить идею его закономерного возникновения.

Многие эволюционисты считают выше своего научного достоинства прятаться за антропный принцип и полагают, что раскрытие законов эволюции должно сделать возникновение разума если не неизбежным, то, по крайней мере, высоковероятным априори. Эту позицию подробно обосновывает известный ученый-популяризатор дарвинизма Ричард Докинз в своей книге «Слепой часовщик». Этим названием Докинз противопоставляет свою книгу трактату богослова Уильяма Пали «Натуральная теология — или признаки и свидетельства существования бога, видимые в явлениях природы» 1802 года, в которой приводится в числе прочих следующий пример. Если бы мы нашли в пустыне часы, то заключили бы, что у этих часов, великолепно выполняющих свое предназначение показывать время благодаря сложности заложенного в их конструкцию замысла, должен быть изготовитель, часовщик. Многие органы человека, такие как глаз, подобны часам и даже гораздо сложнее и более совершенны в своем устройстве, и у них также должен быть создатель. Докинз ставит своей задачей обосновать, что в роли «часовщика», создавшего человека, выступает естественный отбор, который слеп, поскольку у него нет целей, он не знает, что именно создает, он «не смотрит в будущее».

Как же Докинз объясняет создание «слепым часовщиком» сложных адаптационных механизмов живого мира? Для этих целей используется идея нарастающего естественного отбора (градуальности эволюции), заключающаяся в том, что сложный механизм не обязательно должен возникать одновременно. Он может получаться модификацией чуть более простого механизма, а тот, в свою очередь, — модификацией еще чуть более простого механизма. Как говорит Докинз на примере возникновения глаза, «пять процентов зрения — это лучше, чем полное отсутствие зрения», указывая на слова Дарвина: «Если можно было бы продемонстрировать существование какого-нибудь сложного органа, который вряд ли мог быть сформирован множеством небольших последовательных модификаций, то моя теория будет безусловно повержена». Разница между нарастающим отбором и полностью случайным конструированием огромна. Когда нам нужно сделать тысячу шагов в некотором направлении, выбирая каждый раз один из четырех вариантов, тогда, если нам после каждого шага сообщают, в правильном ли направлении мы шагнули, потребуется не более 4000 попыток. Если же нам сообщают правильность результата только после всех шагов, потребуется $4^{1000} \approx 10^{600}$ попыток. Эти цифры просто несоизмеримы! Так, если обезьяна будет случайно печатать на машинке и будут фиксироваться только те ее нажатия, которые удовлетворяют буквам выбранного литературного произведения, то обезьяна действительно сможет напечатать это произведение. Автор приводит множество примеров того, что в природе действует нарастающий отбор, и даже использует интересную модель биоморфов наподобие искусственной жизни для демонстрации этой концепции. Наиболее хорошо следы нарастающего отбора видны в том, что многие новые «изобретения» эволюции надстраиваются над старыми, а не заменяют их (к примеру, низшие животные видят преимущественно движущиеся объекты; у высших же животных для рассматривания статичных сцен сформировался механизм саккад — постоянных быстрых движений глаз).

С Докинзом можно полностью согласиться в его желании опровергнуть распространенное заблуждение, будто дарвиновская эволюция эквивалентна случайному выбору из готовых сложных решений, для которого аргумент исчезающе низкой вероятности удачного выбора абсолютно справедлив. Но Докинз для противопоставления использует другую крайность — нарастающий отбор, который, по сути, сводится к градиентному спуску, т. е. к классическому дарвиновскому «спуску с модификацией». Действительно, в «Слепом часовщике» не раз подчеркивается, что чем больше отличия потомков от родителей, тем хуже работает нарастающий отбор и тем больше он похож на случайный поиск. После знакомства с проблемой комбинаторного взрыва для NP-полных задач должно быть абсолютно очевидно, что градиентный спуск столь же не способен объяснить возникновение разума (или прочих сложных адаптационных механизмов), как и чисто случайный поиск! Конечно, чистому нарастающему отбору потребовались бы какие-то миллиарды шагов для

построения генома человека, и никакие аргументы низкой вероятности ему были бы не страшны (ведь от одноклеточных предков нас отделяют триллионы поколений; кроме того, число одновременно эволюционирующих организмов огромно). Однако градиентный спуск, являющийся жадным алгоритмом, застрянет в первом же локальном экстремуме, и эволюция просто прекратится: любое малое изменение генотипов будет приводить к ухудшению функции приспособленности, которая не может быть монотонной. Если уж при игре в шахматы жадные алгоритмы не приводят к сколько-нибудь интересным результатам, могут ли они привести к ним в случае эволюции?

Докинз сам честно признается, что не понимает, почему в некоторых случаях не произошли «очевидные» эволюционные улучшения, в частности, почему у наутилуса не возник хрусталик, улучшающий зрение благодаря фокусировке света. Если рассматривать эволюцию как поиск, то ответ прост: некоторые эволюционные улучшения требуют преодоления локального комбинаторного взрыва, выполнение чего является весьма маловероятным. Конечно, мы не требуем от эволюции точного решения NP-полных задач, и известно множество несовершенных изобретений эволюции, которые Докинз тоже использует для демонстрации нарастающего отбора, относя к ним, в частности, расположение фоторецепторов на внутренней стороне сетчатки (из-за чего образуется слепое пятно), а также искривление черепа у камбалы (нужное для использования обоих глаз при плавании одним боком по дну). Но если говорить о действительно сложных механизмах и NP-полных задачах высокой размерности, то неоптимальность решения очень быстро превращается в его полную неработоспособность. Даже Докинз на примере простых моделей биоморфов неоднократно наталкивался на явные барьеры на пути дальнейшей эволюции, в результате чего ему вручную приходилось расширять геном своих созданий, руководствуясь знаниями из биологии. Такая ситуация типична для искусственной жизни: поиск в алгоритмически полном пространстве принципиально не может вестись градиентным спуском. Если предположить, что механизмы, принадлежащие алгоритмически полному

пространству, возникли в ходе эволюции, то нельзя согласиться с упрощенной идеей нарастающего отбора и следует признать, что эволюция должна представлять собой гораздо более сложный вид поиска.

Вспомним, что единственная метаэвристика, заимствованная у биологической эволюции в наиболее упрощенном виде, сделала методы эволюционных вычислений достаточно эффективным средством решения проблем поиска (особенно по сравнению с методами как чисто случайного перебора вариантов, так и градиентного спуска) в методах дедуктивного и индуктивного вывода. Какой же эффективностью поиска обладает эволюция в целом?! Ведь в биологической эволюции имеется множество других метаэвристик, которые не моделируются в ГА. К примеру, мутации обладают гораздо более сложным характером, чем это представляется в методах эволюционных вычислений. Помимо замены одной пары оснований на другую может происходить вставка или потеря оснований (при этом, если будет вставлено одно или два основания, то сместится рамка считывания триплетов, что приведет к изменению всех аминокислот). При некоторых мутациях возможны перестройки на уровне хромосом. Все это может иметь большое значение для изменения длины генома (которая в ГА обычно считается фиксированной).

Мы упоминали о существовании большого количества последовательностей ДНК, не кодирующих фенотипические признаки, а управляющих экспрессией других генов. Эволюционный перебор этих последовательностей создает дополнительный уровень оптимизации (оптимизации процесса оптимизации), никак не моделируемый в эволюционных вычислениях. Также в природе и скорость мутации разных генов различна (и различна вероятность их мутации в разных направлениях) и зависит от многих факторов, в частности гены, изменение которых мало сказывается на значении фитнесс-функции, мутируют быстрее. Иными словами, применение самих «генетических операторов» в естественной эволюции выполняется адаптивно, и сами методы адаптивного применения этих операторов, закодированные в ДНК, эволюционируют. Даже Докинз отмечает (но без объяснения в рамках концепции нарастающего отбора), что разные

виды эмбриогенеза имеют разную эффективность не только в смысле выживания, но и в смысле перспективности для эволюции, в связи с чем должна иметь место прогрессивная эволюция способности к эволюции. Эволюция способности к эволюции была даже проверена учеными на практике в долгосрочных эволюционных экспериментах с бактериями (как оказалось, каким-то образом закрепляются не только мутации, дающие мгновенный выигрыш в приспособленности, но и мутации, повышающие эффективность дальнейшей эволюции). Можно сказать, что «слепой часовщик» создавал не только часы, но и очки для себя, постепенно в этом процессе приобретя зрение.

Еще одним примером метаэвристик может служить то, что идеи ближнего и дальнего родства реализуются в эволюции одновременно: скрещивание возможно только между представителями одного вида (ближнее родство), но внутри одного вида из-за существования летальных и полулетальных генов скрещивание близких родственников зачастую приводит к порождению потомков с низким значением фитнесс-функции (дальнее родство). На примере эволюционных вычислений видно, почему это должно быть так. Если представить, что каждый вид располагается на своей моде фитнесс-функции, то межвидовое скрещивание бессмысленно, так как результат попадет между модами. Скрещивание же практически идентичных особей внутри вида также не приводит к оптимизации.

Идея рецессивных и доминантных генов также практически не нашла еще применения в эволюционных вычислениях. А ведь диплоидный набор хромосом позволяет сохранять большое разнообразие генофонда популяции, не вредя при этом фитнесс-функции ее особей, ведь большинство генных мутаций оказываются рецессивными по отношению к немутантному аллелю. Даже при беглом взгляде видна полезная роль этой метаэвристики: она в некотором роде позволяет совместить генетические алгоритмы и поиск с возвратами, характерный для классических методов эвристического программирования. Интересно также, что существуют гены, для которых свойство доминантности и рецессивности не является бинарным (так называемая кодоминантность), т. е.

пара аллелей в гомологичных хромосомах совместно (но, возможно, с разным «весом») проявляется в фенотипе.

Это лишь небольшая часть списка известных на сегодня эволюционных механизмов, часто имеющих вполне отчетливый смысл как эвристик и метаэвристик поиска. На их фоне эволюционные вычисления выглядят игрушечной машинкой по сравнению с настоящим автомобилем. Но самое поразительное то, что все эти эвристики были «изобретены» в ходе самой эволюции! К примеру, первые механизмы рекомбинации ДНК возникли 3,5 миллиарда лет назад, а механизм мейоза с кроссинговером — всего лишь около 850 миллионов лет назад (т. е. на его «изобретение» ушло больше 2,5 миллиардов лет). Регуляторные гены также возникли весьма давно, но далеко не сразу.

Все это показывает, что реальная эволюция — закономерный процесс. Она не похожа на случайный перебор генотипов, в ходе которого благодаря счастливой случайности появился разум. Также она не похожа и на «жадный» градиентный спуск (чистый нарастающий отбор). Объяснение этих закономерностей должно дать понимание истинной причины возникновения разума без привлечения антропного принципа. Для начала эволюцию можно трактовать как самооптимизирующийся поиск. Возможно, этот поиск начинался практически как случайный, однако в ходе него перебирались не только сами объекты, но и эвристики поиска. На обнаружение главных метаэвристик ушли миллиарды лет, но по мере их появления эффективность поиска постепенно возрастала. Такой взгляд на эволюцию, возможно, в чем-то неполон, но позволяет снять (или, вернее, заметно ослабить) проблему низкой вероятности возникновения разума, хотя отнюдь не дает достаточно полного понимания феномена эволюции, а также ставит некоторые новые вопросы.

Для специалистов в области ИИ наиболее интересным является вопрос, насколько исследование эволюции может помочь в решении их проблем. Такое исследование уже позволило получить интересные методы поиска, которые можно расширить при более систематичном анализе эволюционных метаэвристик. Достаточно ясное понимание принципов эволюции должно также давать ответ на вопрос, в какой

мере она может быть воспроизведена на компьютере для построения ИИ.

Пока искусственная эволюция наталкивается на очень большие трудности. Действие чистой «дарвиновской» эволюции (наследственность, изменчивость, отбор), даже дополненной механизмом скрещивания (который, напомним, в естественной эволюции возник далеко не сразу), приводит к ограниченному усложнению поведения. Даже создать неплохо играющую шахматную программу методами эволюционного программирования не получается, что уж говорить о разуме в целом? После знакомства с проблемами решения NP-полных задач этот результат должен выглядеть вполне закономерно, поскольку искусственная эволюция сама собой не превращается в самооптимизирующийся поиск. Можно было бы добавлять некий внешний модуль — супервизор, который будет перебирать разные способы оптимизации эволюционного поиска, делая его направленным в сторону усложнения поведения искусственных организмов. Но такой путь выглядит неестественным. Необходимость введения супервизора связана с тем, что в искусственной эволюции геномы, как правило, не имеют «физической» реализации в виртуальном мире; они вынесены за его пределы вместе с генетическими операторами и даже программами управления («мозгом») искусственных организмов, которые образуют своего рода «духовный мир». При моделировании таким способом нет никакой возможности достижения многих эффектов, к примеру, переноса фрагментов ДНК с вирусами или образования симбиоза между организмами (в том числе на внутриклеточном уровне). Чтобы генетические операторы могли модифицироваться без супервизора, они должны быть частью виртуального мира, который придется моделировать на более низком, физико-химическом, уровне. Но почему физические процессы приводят к возникновению новых эвристик биологической эволюции? Ведь изобретение новых эвристик — очень нетривиальная проблема, которая в области ИИ весьма далека до полного решения. Даже с введением супервизора она остается очень сложной. Как «глупые» физические законы позволяют ее решить? Этот вопрос выводит нас на проблему самоорганизации материи

на более низких, чем биологический, уровнях ее организации. Кроме того, становится понятно, что биологическая эволюция — далеко не самостоятельный феномен.

В эволюции есть следующая странность: наряду со сложными видами существуют и простые, причем простых больше. Это говорит о том, что сложные виды нельзя считать лучше приспособленными, либо о том, что приспособленность не является ключевым фактором эволюции. Ведь, к примеру, полагается, что крокодилы существуют в практически неизменном виде 250 миллионов лет, а латимерии (род кистеперых рыб) — 400 миллионов лет! Почему их не вытеснили более интеллектуальные виды? Мы привыкли к концепции эволюционной лестницы. Но в чем заключается меньшая приспособленность комара по сравнению с шимпанзе или, тем более, с такими вымершими видами, как неандертальцы? Если разум — вершина эволюции, как выжили растения и почему они до сих пор не приобрели черт интеллектуального поведения?

Нередко на эти вопросы можно услышать ответ, что более развитые виды занимают экологические (эволюционные) ниши, которые недоступны более простым видам. Но почему же они не вытесняют эти простые виды из их ниш? То, что идея экологических ниш не дает полного ответа на этот вопрос, демонстрирует следующий пример: люди при заселении в Австралию завезли ряд животных (кроликов, коров, колорадского жука и т. д.), которые там не имели «естественных врагов». Их приспособленность (по крайней мере, локальная) оказалась очень высока, и их быстрое размножение становилось настоящим экологическим бедствием. Человеку приходилось ввозить других животных для того, чтобы вернуть баланс. Это говорит о том, что в любой экосистеме существует множество ниш, для занятия которых вовсе не надо быть высокоразвитым существом. Конечно, интеллектуальное поведение не будет лишним для выживания, но огромное число видов успешно выживают и без него. Почему же в природе эти ниши не заполняются «автоматически», по принципу «выживает сильнейший»? Почему не возникает сверхприспособленных особей типа «суперхищников»?

Отчасти это можно объяснить тем, что оптимизационная задача, решаемая эволюцией, условно говоря, является NP-полной, поэтому возможные усовершенствования происходят вовсе не мгновенно. Они имеют низкую вероятность. Если мы возьмем большое число ниш, то для небольшого их числа эта вероятность реализуется за малый промежуток времени, для большого их числа — за значительный промежуток времени, а для некоторого их числа — не успеет реализоваться за все прошедшее время. В связи с этим наличие «живых ископаемых» или превалирование по численности не особо интеллектуальных видов может вызвать наше удивление, только если мы будем абсолютизировать нарастающий отбор.

Кроме того, в генах заложено много механизмов, *препятствующих* максимальной выживаемости или максимальному размножению. Даже «естественная» смерть является в определенной степени генетически запрограммированной. Откуда природа «знает», что спустя много поколений «суперхищник» вымрет, погубив все живое или отбросив эволюцию на многие миллионы лет назад? Понятно, что эти гены могли возникнуть как нейтральные мутации и присутствовать в генофонде некоторых видов, и постепенно те виды, которые этими генами не обладали, вымерли после своего краткого «триумфа». Среди эволюционистов существует даже отдельное течение, называющее себя «нейтралистами». Они полагают, что большинство мутаций являются нейтральными (для выживаемости), и именно накопление нейтральных мутаций делает возможными большие эволюционные изменения. Мы бы могли провести следующую аналогию: положительные и отрицательные мутации сходны с условным рефлексом, сопровождающимся подкреплением или наказанием, тогда как нейтральные мутации сродни латентному обучению. Именно латентное обучение позволяет преодолеть локальность условного рефлекса, хотя и требует для этого сложных механизмов. Также и нейтральные мутации могли бы стать важной частью мощного эволюционного поиска, если бы имелись механизмы их правильного накопления и использования (к примеру, в форме «мусорной» ДНК). Однако с какой стати сами эти механизмы возникнут, если они

лишь обеспечивают появление генов, которые потенциально обеспечивают выживание видов на большом временно́м масштабе? Возможно, из-за этого первые метаэвристики и формировались миллиарды лет, но даже за счет больших временны́х масштабов нарастающий отбор плохо объясняет их возникновение.

Недостаточность локального действия принципа естественного отбора видна и на следующей проблеме. Многие виды вымерли из-за глобальной катастрофы или изменения климата. Но почему периодические изменения климата должны приводить к появлению видов, способных к этому изменению адаптироваться, вместо просто циклической смены видов, приспособленных к текущему климату? И такие циклические смены действительно известны. В частности, виды могут успевать отращивать шерсть при приближении к ледниковым периодам и укорачивать ее после них, «отслеживая» среднегодовую температуру. В стратегиях поведения хищников и жертв могут также наблюдаться долговременные периодические колебания. Очень быстро мутируют вирусы и бактерии, «отыскивая» бреши в иммунных системах животных. Если рассматривать этот процесс как обучение, то его скорость сопоставима со скоростью обучения высших животных, однако эти мутации не ведут к существенному усложнению вирусов и тем более к возникновению у них интеллекта. Да и насекомые успевают на генном уровне приспосабливаться к применяемым против них химическим средствам за считаные годы! Как же возникают организмы, которые способны приспосабливаться к изменению климата на негенетическом уровне? Конечно, никто не запрещает случайно появляться таким видам в ходе эволюции. Но поскольку не вполне ясны их преимущества с точки зрения выживаемости, они не будут создавать «вектор эволюционного развития».

Представим, что у нас есть фиксированная функция приспособленности. Эта функция будет детерминированным (хотя и неявным) образом задавать оптимальный вид организмов, что сделает эволюцию весьма рутинной процедурой поиска этого вида с предопределенным результатом. Как уже говорилось, функция приспособленности сама меняется при

появлении новых видов, но чем же определяется то, в какую сторону направлены эти изменения? Естественный отбор принципиально не может дать ответ на этот вопрос. Понятие эволюционно устойчивых стратегий, о котором упоминалось выше, дает чуть лучшее объяснение. Нужно говорить не об эволюции отдельных видов, а об их коэволюции, или даже эволюции биосферы в целом. Однако это не дает ответа на вопрос, откуда в такой эволюции берется направленность на возникновение разума. Идет направленное усложнение систем без видимой реальной причины.

Можно, конечно, сказать, что человек себе льстит в том, будто он является (на настоящий момент) венцом эволюции, и что он — вполне рядовой ее продукт, один из не особо примечательных миллионов видов. Но проблема именно в том, что построение разума — это очень сложная задача, которая не может быть решена случайно, как побочный продукт эволюции.

Возникновение разума может быть объяснено, только если саму эволюцию представить как самооптимизирующийся поиск, *направленный* на построение разума (или, по крайней мере, на усложнение). Возвращаясь к метафоре «мышление как поиск», можно прийти к выводу, что между мышлением и эволюцией имеется подозрительно много общего. По крайней мере, эволюция «научилась» делать то, чего до сих пор так не хватает программам ИИ — изобретать новые эвристики (чуть ли не единственная программа ИИ, при создании которой эта проблема серьезно рассматривалась, — «Эвриско» Дугласа Лената), причем в алгоритмически полном пространстве.

Однако такой взгляд преимущественно чужд эволюционистам. Пытаясь опровергнуть креационизм, они пытаются избегнуть каких-либо параллелей с работой мышления. При этом они ограничиваются описанием эволюционных механизмов (часть из них выше была упомянута), которые в совокупности позволили бы сделать естественное возникновение разума человека хоть немного вероятным. Однако этого недостаточно. Представим себе опять шахматы. Мы смотрим на фигуры и считаем, что они движутся по физическим законам: пешки движутся только прямо, слоны —

по диагоналям и т. д., объясняя детали случайностью. Но мы замечаем странность: совокупное движение фигур отвечает решению сложной проблемы — хорошей игре, что при случайном выборе ходов маловероятно. Именно это мы наблюдаем в случае эволюции. Мы начинаем уточнять конкретные механизмы того, как ходят фигуры. Обычно они стремятся съесть друг друга, занять положение типа вилки и т. д. Также и при описании эволюции мы указываем конкретные механизмы, которые частично объясняют ее эффективность. Но чем больше мы находим и описываем эвристик, тем больше это описание напоминает не столько совокупность законов движения фигур, сколько интеллектуальную программу, играющую в шахматы. То же самое и с эволюцией. Чтобы избавиться от возражений креационизма, нужно описать не текущую «интеллектуальную программу эволюции», а то, как она могла возникнуть.

Конечно, эволюция обладает некоторыми несомненными атрибутами интеллектуального, самообучающегося процесса. Но при признании параллелей между мышлением и эволюцией следует предостеречь от поспешного переноса на нее всех прочих свойств человеческого разума (таких как, например, самосознание или целеполагание). Вряд ли стоит думать, что за эволюцией стоит мощный разум, существовавший с начала времен. Иначе этому разуму не понадобились «размышления» на протяжении нескольких миллиардов лет, за которые он «придумал» человека (по крайней мере, человек надеется построить ИИ гораздо быстрее). Скорее, эволюция в начальный момент больше похожа на младенца, обладающего самым минимальным интеллектом, но способного к обучению (исходно на основе обширного стохастического поиска).

Интересно, что в настоящее время существуют попытки создания моделей самооптимизирующегося искусственного интеллекта, который бы был максимально универсальным (т. е. в него был бы заложен минимум априорной информации при выполнении свойства алгоритмической полноты). Как правило, эти модели являются расширением универсальной модели предсказания на основе алгоритмической вероятности, которая применяется для выбора наилучших

действий при обучении с подкреплением. При этом для них доказываются различные свойства оптимальности. Несмотря на безусловную теоретическую значимость, такие модели неприменимы на практике, в частности потому, что они требуют от «универсального интеллекта» накопления в ходе жизни всей информации, которая в естественный интеллект заложена эволюционно (включая и эвристики эффективной самооптимизации). Весьма интересно, что подобные модели максимально простого самооптимизирующегося интеллекта больше напоминают именно эволюцию, а не естественный интеллект человека и животных.

Даже если признать эволюцию таким самообучающимся процессом, возникает вопрос, как он реализован (ведь у эволюции нет никакой централизованной системы управления). Одной идеи естественного отбора недостаточно. Повторим еще раз, что модели искусственной жизни, включающие «естественный» отбор, не дают такого повышения сложности поведения организмов, как в природной эволюции, даже если явно заложить в эти модели цель создания таких организмов — слишком высока сложность данной задачи. Необходима оптимизация самого процесса поиска, который в моделях ИЖ обычно фиксирован и вынесен за пределы моделируемого мира. Для воспроизведения появления новых метаэвристик в искусственной эволюции требуется помещение генов внутрь мира, их физическое моделирование. Так же и глубинные механизмы биологической эволюции имеют физическую основу.

Ограниченность эволюционных вычислений можно продемонстрировать на примере невозможности (без детального моделирования физических процессов), с их помощью изобрести новые типы сенсоров, что является несравненно более простой задачей, чем «изобретение», скажем, мейоза или регуляторных генов. В то же время простейшие физические реализации генетических алгоритмов могут давать совершенно неожиданные результаты. Интересный случай произошел в Университете Сассекса, где ГА применялись для оптимизации осцилляторных схем, воплощавшихся физически. К удивлению исследователей, в результате «эволюции» возник примитивный радиоприемник, регистрировавший

сигнал от стоящего рядом оборудования, хотя такая возможность исходно совершенно не предусматривалась. Конечно, обеспечить полноценную «воплощенную» эволюцию интеллектуальных систем крайне проблематично. Вместо этого нужно рассмотреть вопрос о том, какие базовые физические принципы следует промоделировать для организации неограниченной виртуальной эволюции. Возникает принципиальный вопрос, могут ли физические процессы объяснить самоорганизацию материи для начала биологической эволюции и ее поддержания в смысле порождения новых метаэвристик.

САМООРГАНИЗАЦИЯ

ИМИТАЦИЯ ОТЖИГА

Может ли мертвая, «косная» материя обладать свойствами, интересными для области ИИ? Мы уже упоминали об оптических вычислениях и квантовых компьютерах, правда, они предназначены для решения проблем быстродействия, а не структуры информационных процессов. Этого недостаточно для объяснения глубинных механизмов эволюции. Можно было бы списать загадки возникновения жизни, начала биологической эволюции на какие-то неалгоритмические процессы. Однако это делать рано: физические процессы оказываются интересными и с алгоритмической точки зрения. В частности, модели таких физических процессов, как рост кристаллов, оказались весьма востребованными в ИИ при решении задач поиска и оптимизации, т. е. именно тех задач, с которыми связана эволюция.

Кристаллы — удивительная вещь. Непросто вырастить монокристалл сахара или соли в домашних условиях или алмазы промышленно. Даже если это удается, человек лишь создает условия, а растут кристаллы сами. Интересным их свойством является минимум потенциальной энергии атомов. Еще интереснее то, что этот минимум достигается не всегда, а только при относительно медленном выращивании.

сих пор представляющей далеко не только исторический интерес. Необходимо отметить, что, несмотря на недостаточную проработанность механизмов метасистемных переходов, сама их концепция легла в основу идеи суперкомпиляции в области языков программирования, которая привела к созданию одного из первых функциональных (и в то же время декларативных) языков программирования — РЕФАЛ (РЕкурсивных Функций АЛгоритмический). Можно сказать, что полноценная теория метасистемных переходов сама могла бы стать основой нового глобального метасистемного перехода, и применение такой теории к алгоритмическим системам представляет несомненный интерес. Закономерность совершения предыдущих метасистемных переходов вполне естественно наводит на мысль об их продолжении в будущем.

ТЕХНОЛОГИЧЕСКАЯ СИНГУЛЯРНОСТЬ

Мало кто способен отрицать взрывное развитие техники. Если обычная продукция дорожает из-за инфляции, то многие виды высокотехнологичной продукции преодолевают эту тенденцию и дешевеют, при этом характеристики их улучшаются. Для классической экономики это парадоксально. Представьте себе фермера, который выращивает все более хорошую картошку и продает ее все дешевле!

Многие виды высокотехнологичной продукции для большинства людей появляются совершенно неожиданно, поначалу являясь предметом роскоши, но постепенно становясь общедоступными. Так было с электрическими лампочками, телевизорами, мобильными телефонами. Так происходит с бытовыми роботами. Технические новинки все быстрее появляются и все быстрее устаревают. Их полные возможности уже нередко игнорируются многими взрослыми как необязательные для жизни и осваиваются лишь детьми. Можно было бы привести много примеров, но они очень быстро устареют. Дальнейшее увеличение темпов развития техники приведет к тому, что люди просто не будут успевать ею овладевать и она просто не будет востребована.

Наиболее ярко развитие техники проявляется на примере компьютеров. Широко известен закон Мура об удвоении мощности компьютеров (числа транзисторов на кристалле) раз в 18–24 месяца, предложенный в 1965 году. О том, как долго этот закон будет выполняться (и выполняется ли сейчас), до сих пор идут споры. Нередко говорится о физических ограничениях на скорость распространения сигналов, размеры элементов, энерговыделение из-за необратимости вычислений и т. д. Производители идут на разные хитрости для того, чтобы поддержать выполнение этого закона, например на создание многопроцессорных систем. Производительность наиболее мощных суперкомпьютеров продолжает удваиваться примерно раз в 1,5 года. Ведь нарушение этого закона будет рассматриваться как прекращение прогресса в компьютерной технике, которое считается нежелательным, несмотря на то, что из накопленных вычислительных мощностей людьми используется сейчас ничтожная часть.

Сходная ситуация еще раньше проявилась в науке. Рост числа научных журналов, конференций и публикуемых статей экспоненциальный. Ученые давно не успевают прочитать не просто все научные статьи, но даже статьи в своей области. Как определить, какие из множества статей стоит изучить? Можно опираться лишь на мнение других ученых. Как результат, научное сообщество пытается как-то «самоорганизоваться». В частности, стал широко использоваться такой показатель, как индекс цитирования, определяющий так называемый импакт-фактор журналов. Казалось бы, журналы, статьи из которых много цитируются, должны содержать лучшие материалы. Но посмотрите: ученые будут цитировать лишь те статьи, которые прочитали, а читать они будут статьи в журналах с высоким импакт-фактором. Получается система с положительной обратной связью: чем больше у журнала импакт-фактор, тем больше его статьи цитируют, и наоборот. Кроме того, чем больше у журнала импакт-фактор, тем больше авторов там захочет опубликовать свои работы. Само качество статей здесь как будто ни при чем. Импакт-фактор журналов сродни высоте деревьев в лесу. Конечно, в журналы, в которые поступает большее количество статей, могут отбираться лучшие из них. Но

как определить лучшие статьи до того, как с ними познакомится научное сообщество (которое как раз не успевает в равной мере познакомиться со всеми статьями)? Выше мы видели не один пример, когда статьям с фундаментальными открытиями было отказано в опубликовании в рецензируемых журналах. Выходит, в журналах с высоким импакт-фактором будет отдаваться предпочтение статьям, которые будут понятны максимально широкой аудитории ученых. Таким образом, проблема эффективного распространения действительно важных новых научных знаний остается нерешенной и все более усугубляется.

Складывается впечатление, что ограниченные возможности самого человека скоро станут (если уже не стали) основным ограничением научно-технического прогресса (хотя полезная занятость мозгов людей столь же низка, как и компьютеров). Окажется ли этот прогресс кратковременной вспышкой, быстро возникшей и также быстро потухшей? На этот вопрос можно было бы дать положительный ответ, если бы текущее развитие техники не было лишь одним из многих подобных событий в эволюции.

Так, иногда возникают сомнения, что человеческий разум произошел от обезьяньего. Ведь до этого эволюция шла миллиарды лет, и животные умнели как будто медленно. А потом за «ничтожное» время — в тысячи раз меньшее длительности предыдущей эволюции — произошел настоящий «взрыв» разума. Почему он вдруг произошел? Да и мог ли он вообще произойти? Но обратите внимание: «взрыв» технологий имеет сходные характеристики. Его продолжительность составляет тысячные доли от длительности существования человечества.

Экспоненциальный рост — достаточно заурядное событие как в научно-техническом, так и в эволюционном развитии. Возрастание сложности генотипов в процессе эволюции, видимо, можно считать экспоненциальным, но период удвоения их сложности составлял около сотни миллионов лет. После того как возникла нервная система, сложность ее также стала возрастать экспоненциально с периодом удвоения в десятки миллионов лет. На этом фоне возникновение человеческого разума — лишь один из этапов, характеризующийся более коротким периодом удвоения сложности, чем предыдущие этапы.

Также и отличие закона Мура лишь в том, что время удвоения производительности компьютеров является много меньше времени жизни человека, поэтому воспринимается им как весьма короткое. И до этого происходили события, сопровождающиеся экспоненциальным ростом сложности информационных систем. В качестве таких событий можно назвать изобретение письменности, а позднее — книгопечатания, за которыми следовал экспоненциальный рост количества сохраняющейся в данной форме информации с периодами удвоения в сотни и десятки лет соответственно. Да и после компьютерного «взрыва» уже появлялись системы с экспоненциальным ростом сложности. К таким системам относится Интернет, период удвоения сложности которого составил всего несколько месяцев.

Экспоненциальный рост числа элементов в этих системах похож на «разрастание предпоследнего уровня» перед метасистемным переходом, после которого появляется новый системный уровень. Кстати, экспоненциальное развитие науки отмечал и Турчин. Но в дополнение к самому факту метасистемных переходов здесь в глаза бросаются динамика и временны́е характеристики этих процессов. Помимо того, что на каждом уровне происходит экспоненциальное возрастание сложности, период удвоения после каждого метасистемного перехода сокращается, а также сокращается время между последующими переходами.

Эта ситуация очень похожа на сокращение периода между бифуркациями на диаграмме логистического отображения, рассмотренного выше. За конечное время количество бифуркаций у этого отображения оказывается бесконечным. Подобное поведение известно для систем с нелинейной (самоусиливающейся) положительной обратной связью и носит название режимов с обострением. Также и число метасистемных переходов может оказаться бесконечным за конечное время. Гипотетическая точка, в которой сложность информационных систем оказывается бесконечной, получила название *технологической сингулярности*.

Идею о достижении техническими системами бесконечной сложности за конечное время высказал Вернон Виндж

в начале 90-х годов прошлого века. Исходно эта идея иллюстрировалась примерно следующим рассуждением. Пусть быстродействие компьютеров удваивается раз в 16 месяцев. Представим себе искусственный разум человеческого уровня. Через 16 месяцев этот разум будет перенесен на процессоры, в два раза более быстрые, в связи с чем сможет разработать процессоры нового поколения в два раза быстрее, т. е. за 8 месяцев. Через 8 месяцев он будет мыслить в два раза быстрее и разработает следующее поколение процессоров еще в два раза быстрее, т. е. за 4 месяца. И так далее. Менее чем через еще один год производительность процессоров устремится в бесконечность.

Это рассуждение уязвимо для критики: узким местом для увеличения темпов удвоения мощности процессоров станет не скорость их разработки, а скорость совершенствования технологической базы (построение фабрик для реализации новых технологий). Еще ярче эта проблема видна на другой самоускоряющейся технологии — генной инженерии. Гипотетически возможно такое усовершенствование генома человека, которое приведет к усилению его мыслительных процессов. Более умные люди смогут еще сильнее и быстрее улучшить геном следующего поколения. И так далее. Несмотря на это, бесконечного ускорения развития достигнуть не удастся из-за ограниченной скорости эмбрионального развития: следующее поколение не сможет появляться мгновенно. Однако сама идея нелинейной положительной обратной связи здесь проиллюстрирована весьма точно. Требуется лишь уточнить, что увеличение сложности информационных систем должно происходить не путем простого увеличения быстродействия компьютеров (или человеческого мозга), а путем цепочки метасистемных переходов со сменой типа прогрессирующих информационных систем.

Сейчас концепция технологической сингулярности рассматривается не просто как результат локального технического прогресса, но как результат всей предыдущей эволюции, для которой можно построить кривую возрастания сложности информационных систем. Разные авторы немного по-разному строят эту кривую в зависимости от выбора того параметра, который отражает сложность систем, а также

от того, какие метасистемные переходы принимаются за основные. Легче всего оценивается информационная емкость систем, в связи с чем рассматривается смена носителей информации: гены, нервная система, книги, компьютеры, для которых мы уже отмечали характерные времена удвоения сложности. Конечно, для каждого из этих носителей имело место множество промежуточных метасистемных переходов разной значимости.

При появлении нового типа носителей информации уже не так важно, продолжается ли экспоненциальное усложнение предыдущего уровня. Ведь период удвоения следующего уровня существенно меньше, поэтому он за короткое время обгонит по сложности предыдущий уровень. Так произошло с генами: исходно нервная система была гораздо менее вместительной, чем геном, но относительно быстро емкость мозга превысила емкость генома. Сейчас уже не столь важно (в смысле возрастания сложности), идет ли биологическая эволюция (вернее, накапливается ли новая информация в генах), поскольку скорость развития новых информационных систем несоизмеримо выше. То же произошло и с человеческим мозгом: его емкость, возможно, и продолжает увеличиваться, но это происходит слишком медленно на фоне новых носителей информации. Аналогичное заключение можно сделать и относительно компьютеров: не так важно, продолжит ли выполняться закон Мура, если компьютеры перестанут быть «передним фронтом глобальной эволюции». В этой связи формулируется общий закон возрастания сложности, гласящий, что время удвоения сложности каждого нового типа информационных систем так же, как и время возникновения новых типов систем, сокращается в некоторое число раз.

Интересно, что во всех случаях: генотип, мозг, книги, компьютеры, — до возникновения нового вида информационных носителей наблюдается однотипный сценарий интеграции носителей предыдущего уровня. Когда наибольшей емкостью обладали генетические системы хранения информации, возникновение полового размножения привело к возможности обмена информацией между разными генотипами. Вместо генотипов единой информационной системой стали

генофонды (включающие разные комбинации генов одного вида). При развитии мозга возник язык как способ обмена информацией между разными нервными системами, и единой информационной системой стал социум. Письменность, позволившая информацию, накапливаемую нервными системами, переводить в более долговечную форму, дополнилась книгопечатаньем (хотя, возможно, появление библиотек, содержащих в себе разные книги, является более корректной иллюстрацией к интеграции информации на данном уровне). Очевидно, роль Интернета для компьютеров такая же, как и роль скрещивания для генов или языка для отдельных нервных систем. Таким образом, исчерпание возможности экспоненциального роста производительности компьютеров вовсе не будет являться преградой на пути к следующему глобальному метасистемному переходу. Вероятнее, это будет индикатором скорого появления принципиально нового «переднего фронта эволюции».

Новый метасистемный переход — это еще не сингулярность. Сингулярность возникает в результате каскада переходов со все уменьшающимся периодом. Однако о ближайшем переходе можно высказать хоть какие-то предположения. Футурологи предлагают несколько основных сценариев возникновения сингулярности, связанных с так называемыми сингулярными технологиями, к которым относят самоприменимые технологии типа информационных технологий (особенно искусственного интеллекта), биотехнологий (особенно генной инженерии), а также нанотехнологий (особенно наноассемблеров). Эти технологии могут участвовать в подготовке метасистемного перехода как независимо (например, в форме возникновения автономного искусственного разума), так и совместно (например, в форме считывания структуры человеческого мозга с помощью нанороботов и загрузки его в компьютер с последующим возникновением коллективного сознания на основе сети Интернет). Во всех этих сценариях, правда, дальнейшая эволюция так или иначе связана с развитием интеллекта. В своих сценариях футурологи, конечно, выходят далеко за рамки науки — в область фантазии (высказываются даже мысли о непосредственном объединении сознаний людей, число которых приближается

к числу клеток мозга, в связи с чем иногда вспоминается так называемый резонанс Шумана, связанный со стоячими электромагнитными волнами между поверхностью и ионосферой Земли, частота которых близка к альфа-ритму человеческого мозга). Однако реальность, вероятно, окажется иной и превзойдет самое смелое воображение, которому сейчас для предсказания не хватает знаний.

В частности, футурологами, как и фантастами, искусственный интеллект обычно изображается очень похожим на человеческий. Конечно, всегда вводятся какие-то отличия. ИИ обычно показывают менее творческим, эмоциональным и т. д., но более логичным, хорошо считающим, с лучшей памятью. Эти различия (которые могут оказаться неверными для настоящего ИИ) лишь подчеркивают то, что человеку ИИ удается представить как немного искаженный образ своего собственного мышления. В то же время, вероятно, развитие интеллекта будет сопровождаться далеко не только улучшением аппаратных характеристик.

Поскольку мы с трудом представляем себе следующий метасистемный переход, предсказать, какую форму примет каскад таких переходов, просто невозможно. Но можно ли, построив кривую возрастания сложности, предсказать хотя бы примерный момент наступления сингулярности? Эта оценка в зависимости от способа ее выполнения несколько отличается у разных исследователей. Наиболее «оптимистично» настроенные авторы приводят оценки, соответствующие примерно 2020 году. Более осторожные авторы относят момент наступления сингулярности вплоть до 2050 года. Мало кто из авторов, верящих в реальность сингулярности, относит ее на заметно более поздние годы. Сейчас наступление столь глобального события в столь короткие сроки звучит фантастично, но, вероятно, не более фантастично, чем современность представилась бы нашим предкам. Кроме того, человеческий разум просто не привык оперировать даже экспоненциальными зависимостями (при обсуждении NP-полных задач уже приводился пример о листе бумаги, сложенном пополам 50 раз подряд). Мы здесь не будем углубляться в эти оценки, которые порой бывают спорными. Важен сам факт того, что промежутки времени

между метасистемными переходами сокращаются, и если эта тенденция сохранится, технологическая сингулярность наступит в обозримое время.

Однако именно вопрос о сохранении тенденций является наиболее спорным. С одной стороны, отрицание сингулярности в основном опирается на наличие чисто физических ограничений на максимальную производительность информационных систем. Но эти оценки говорят лишь о сомнительности возможности достижения истинно бесконечной сложности. С другой стороны, отсутствуют содержательные модели, обосновывающие неизбежность последующих метасистемных переходов. В этой связи можно полагаться лишь на эмпирический закон возрастания сложности. Из индуктивного вывода мы знаем, что продолжение любой последовательности нельзя предсказать однозначно: могут быть выдвинуты разные гипотезы. Сингулярность (как следствие сокращения периода между последующими переходами) — наиболее простая (по форме графика) и вероятная гипотеза, но это не значит, что она окажется истинной. Сложность модели, в которой промежутки времени между последующими метасистемными переходами начнут увеличиваться, и кривая сложности симметрично изогнется, лишь немного выше, т. е. ее вероятность незначительно меньше. Конечно, и в этом случае изгиб кривой возрастания сложности будет являться особой точкой, знаменующей изменение характера всего эволюционного развития, но такая особая точка для человечества будет не столь заметна.

С последующими метасистемными переходами (даже без наступления сингулярности) связана опасность того, что человек перестанет принимать участие в дальнейшем развитии. Так, раз возникнув, ИИ человеческого уровня очень быстро станет несопоставимо умнее человека. Другие сценарии также несут опасность потери человеком контроля над дальнейшим развитием, что вряд ли может считаться желательным. Однако официальный запрет на разработки в сфере всех сингулярных технологий (что имеет место в действительности по отношению к клонированию человека в некоторых странах) и тем более разгром отдельных лабораторий (как это иногда представляется в фантастике) вряд ли может остановить раз-

витие и даже просто существенно повлиять на его динамику. Глобальность закона возрастания сложности, действие которого началось задолго до появления человека, заставляет поверить в объективность процессов повышения сложности систем и неизбежность последующих метасистемных переходов вне зависимости от воли человека. Конечно, сейчас именно люди реализуют эти переходы. Так, компьютеры были изобретены вполне конкретными учеными. Упрощенно говоря, практическая неизбежность переходов связана с тем, что если изобретение не будет сделано одними исследователями, то будет сделано другими. На что могут повлиять люди, так это на то, во что воплотятся последующие переходы. Ведь любые самоусиливающиеся процессы, особенно режимы с обострением, характеризуются крайней нестабильностью, возможностью разрушения системы. В этой связи к исследованиям в области ИИ необходимо отнестись со всей серьезностью. Не удивительно, что упоминавшаяся концепция дружественного ИИ разрабатывается именно в Институте сингулярности искусственного интеллекта (Singularity Institute for Artificial Intelligence), созданном на рубеже тысячелетий с целью снижения соответствующих рисков.

Даже если истинная сингулярность не наступит, а связанные с нею риски преувеличены, закон возрастания сложности представляет несомненный интерес. Если сам факт существования человеческого разума можно было списать на естественный отбор, антропный принцип или акт творения, то для объяснения конкретной закономерности возрастания сложности этих концепций уже недостаточно. Этот закон однозначно свидетельствует о том, что наш разум возник как промежуточная стадия неких глобальных процессов, которые можно интерпретировать как самооптимизирующийся поиск. Само человеческое мышление является не только продуктом, но и средством осуществления этого поиска, точнее, одним из его уровней. Интересна обратная связь между уровнями: возможность сознательного улучшения генома аналогична обсуждавшемуся эффекту адаптивного резонанса в восприятии и мышлении. Человека можно было бы считать очередным изобретением эволюции, позволяющим эффективнее вести поиск в пространстве геномов. Од-

нако сама биологическая эволюция является лишь одной, хотя и большой, эпохой в глобальной эволюции, которая включает самоорганизацию на физическом и химическом уровнях, а также технический прогресс. В этой связи не исключена возможность распространения «эволюционного адаптивного резонанса» и на более глубокие уровни (вплоть до оптимизации самих физических законов), но здесь мы опять входим в область чистых догадок.

Кривую возрастания сложности можно пытаться продолжить не только в будущее, но и в прошлое. Вселенной потребовалось свыше 10 миллиардов лет, чтобы появилась Земля с зародившейся на ней жизнью (мы, правда, не знаем, не возникала ли она где-то раньше). Сложность начальных репликаторов оценивается в сотни бит. Десяти периодов удвоения сложности продолжительностью в миллиард лет каждый хватило бы для возникновения этих репликаторов. Такой период удвоения сложности молекул в несколько раз больше периода удвоения сложности генотипов, что вполне соответствует характеристикам предыдущего уровня. Мы упоминали и о возможных промежуточных метасистемных переходах, отделяющих обычные молекулы от современных ДНК-репликаторов. Однако образование атомов и простейших молекул, информационная емкость которых может быть определена по числу состояний, в которых они могут находиться, возросла очень быстро после Большого взрыва, в связи с чем некоторые исследователи полагают, что наша Вселенная образовалась с некоторой «априорной информацией» об этих системах, и привлекают идею множественности вселенных (мультиверса) для объяснения источника этой информации. В такой модели начальный слепой перебор вариантов выполняется на уровне вселенных с разными физическими законами, выступающих своего рода генами вселенных.

Эта гипотеза совместима с некоторыми физическими теориями, в частности, с М-теорией, являющейся обобщением разных версий теории суперструн. Также гипотеза мультиверса связана с многомировой интерпретацией квантовой механики Эверетта, на основе которой Д. Дойч развивает идею квантовых компьютеров (что описано в его книге «Структура реальности»). Интуитивно эта гипотеза кажется

привлекательной, поскольку «объясняет» происхождение значений физических констант, да и типов взаимодействия в нашей Вселенной. Как еще можно ответить на вопрос, почему в нашей Вселенной именно такие, а не другие, физические законы? И почему эти законы так точно подобраны для обеспечения возможности возникновения жизни? Ведь, как предполагается, незначительное изменение некоторых констант привело бы, например, к невозможности термоядерного горения гелия в звездах, в результате чего не образовывались бы более тяжелые химические элементы и жизнь стала бы вряд ли возможна. Идея мультиверса хоть и не имеет на данный момент проверяемых следствий (т. е. остается вне научных рамок), дает непротиворечивый ответ на поставленные вопросы.

Концепция мультиверса также дает свой ответ на следующую проблему. Слишком многие данные наблюдений свидетельствуют о том, что наша Вселенная имела некоторое начало. Сюда относится, в первую очередь, множество астрономических данных, подтверждающих теорию Большого взрыва. Такие феномены, как реликтовое излучение или увеличение красного смещения галактик с расстоянием, указывают на расширение Вселенной. Более того, в звездах преимущественно «выгорают» легкие химические элементы и не наблюдается процессов, в которых бы эти элементы восстанавливались, что свидетельствует о необратимом изменении химического состава Вселенной. С учетом конечной скорости этого изменения Вселенная в известном нам виде не могла существовать бесконечно долго, т. е. она должна иметь начало. Однако очень сложно себе представить, что до возникновения Вселенной ничего не было. Ведь если не было ничего, то не было и времени. Человек же может представить себе возникновение чего-либо только как событие во времени, в связи с чем возникновение самого времени выглядит парадоксальным (хотя может просто свидетельствовать об ограниченности человеческого мышления). Гипотеза мультиверса устраняет этот парадокс, постулируя существование некоторого вечного пространства, в котором в произвольные моменты времени возникают вселенные наподобие той, в которой мы живем. Есть и другое решение этого парадокса в

рамках модели единственной Вселенной, которая периодически (бесконечно длительное время) претерпевает стадии разлета после Больших взрывов и последующего сжатия.

Закон возрастания сложности усугубляет эту проблему. Продолжение этого закона в прошлое и будущее упирается в две крайние точки — Большой взрыв (точнее, состояние с минимальной сложностью информационных систем) и технологическую сингулярность. В отличие от гипотетического гравитационного коллапса Вселенной, технологическая сингулярность не подразумевает бесконечной смены циклов рождения и гибели Вселенной. Возможное бесконечное существование мультиверса плохо укладывается и в конечные временны́е рамки наступления сингулярности.

Концепция развития вообще плохо уживается с идеей отсутствия парадоксального «начала времени». Ведь и у Гегеля (введшего эту концепцию в философию) развитие «мирового разума» имеет начало и конец. Последний связывается философом с осознанием «мировым разумом» себя через учение самого Гегеля (хотя, если уж и трактовать глобальную эволюцию как процесс познания «мировым разумом» себя, то окончание этого процесса нужно связать с технологической сингулярностью). Конечно, кривую возрастания сложности можно продолжить и на любое время назад: если считать, что сложность систем может быть любым дробным числом меньше одного бита, то эта кривая будет просто асимптотически приближаться к нулю при времени, стремящемся к минус бесконечности (что, правда, выглядит немного надуманно). В противном случае, нужно признать либо то, что вся реальность действительно существует конечное время, либо то, что возрастание сложности является локальным по времени, и никакого глобального развития на большем интервале времени в действительности нет (к примеру, вся информация, накопленная Вселенной в текущем цикле, уничтожается при переходе на следующий цикл).

Не исключено также, что все это не более чем метафора. Создание искусственного интеллекта, являющегося одной из наиболее вероятных следующих ступеней на пути глобального эволюционного процесса, возможно, позволит хоть немного приблизиться к ответу на эти вечные вопросы.

ЗАКЛЮЧЕНИЕ

В этой книге мы попытались рассмотреть основные проблемы и достижения в области искусственного интеллекта. Попробуем теперь ответить на вопрос, возможно ли создание сильного ИИ и чего для этого не хватает. Если не накладывать никаких ограничений на способы его создания, то под искусственным разумом можно понимать и искусственно сконструированный мозг, состоящий из биологических нейронов. Возможность построения такого ИИ можно отрицать, пожалуй, только с религиозной позиции, основанной на вере. В основном неверие людей в реализуемость хоть в какой-то форме ИИ связано с подсознательной попыткой защитить свою уникальность, которой исследования мышления как будто угрожают. Чтобы не углубляться в неуместную дискуссию на этот счет, процитируем достаточно стандартные (для специалистов, изучающих мышление) высказывания на этот счет:

«У некоторых людей может возникнуть опасение, что такого рода материалистическая концепция, рассматривающая мозг как некую супермашину, лишит нашу жизнь очарования и отвратит нас от духовных ценностей. Это сродни опасению, что знание анатомии человека помешает нам восхищаться формами человеческого тела. Художники и медики знают, что верно как раз обратное». *(Д. Хьюбел)*

«Современным исследователям мышления уличный „здравый смысл” приписывает стремление все на свете формализовать и обессмыслить творчество через создание механических суррогатов мысли. Реально происходит обратное. Каждая программа искусственного интеллекта — это метафора мысли, требующая для своей реализации подъема на новую ступень мышления, разрушающая стереотипическое мышление „одномерного человека”, не ведающего, что он творит». *(М. В. Сергеев)*

Нас, однако, интересовал более конкретный вопрос о реализуемости алгоритмического ИИ, против которой выдвигаются гораздо более наукообразные аргументы. Эти аргументы можно в основном разделить на математические (связанные с алгоритмической неразрешимостью и проблемами порождения новой информации), физические (отталкивающиеся от тезиса, что мышление или его часть — это специфический физический процесс, который нельзя заменить никаким процессом другой природы) и психологические (указывающие на то, что с помощью алгоритмов не могут быть реализованы такие когнитивные функции, как понимание, переживание эмоций, самосознание и т. д.).

Однако, как мы видели, многие аргументы перестают работать применительно к открытым безостановочным алгоритмическим системам (или воплощенному интеллекту). В частности, для таких систем некорректным оказывается доказательство неразрешимости проблемы останова (и множество следствий из него), а также неприменимыми оказываются теоремы Гёделя, накладывающие ограничения на возможности формальных систем. Обоснования недостижимости функции понимания, отчасти справедливые для замкнутых систем, оказываются некорректными по отношению к системам, осуществляющим формирование понятий в результате взаимодействия с реальным миром. Часто у человека внутренний протест вызывают высказывания о том, что можно запрограммировать мораль и нравственность, но для таких обучаемых систем (снабженных также эмоциональной подсистемой) корректнее будет говорить об их «воспитании», а не программировании. Физические же аргументы, не подкрепленные указанием на то, какие именно функции не могут быть выполнены алгоритмическими системами (будь то решение неразрешимых задач или понимание), оказываются весьма слабыми: в конце концов, если какие-то иные процессы обеспечат искусственному интеллекту то же поведение, которое демонстрирует человек, нам будет не важно, являются ли эти процессы «мышлением» в таком узкофизическом смысле.

Иногда в качестве аргументов против алгоритмического подхода приводятся архитектурные различия между компьютерами и мозгом, которые как будто свидетельствуют о невозможности алгоритмического воспроизведения упоминавшихся когнитивных феноменов. Однако, как писал ведущий специалист в когнитивной психологии Б. М. Величковский, «…Исследование

этих феноменов обычно ведется в контексте дихотомических теоретических противопоставлений: последовательный или параллельный, периферический или центральный, непрерывный или дискретный, врожденный или приобретенный… При этом действительно важные вопросы остаются неизученными». Под действительно важными вопросами понимается структура процессов, реализующих соответствующие операции. И мы действительно неоднократно это видели. Так, дискретные операции могут реализовываться на основе непрерывных процессов, и наоборот (если, конечно, речь не идет о нереалистичных бесконечных точностях), а врожденные навыки формируются в процессе эволюции, которая по структуре немногим отличается от обучения. Для описания структуры процессов ничего лучше алгоритмов еще не придумано.

Пока рано говорить о том, что все загадочные феномены — от самоорганизации до самосознания — могут быть описаны в алгоритмических терминах. Однако, по крайней мере, именно вопрос о том, что эти феномены представляют собой с алгоритмической точки зрения, позволяет избавиться от всякой мистики и гораздо точнее сформулировать их сущность.

Таким образом, серьезных научных аргументов против возможности алгоритмического ИИ в настоящее время нет, а имеющиеся аргументы порой даже противоречат друг другу. В частности, можно одновременно встретить утверждения о том, что человек может решать NP-полные задачи, что недоступно компьютеру (а значит, подразумевается гарантированное нахождение полного решения), и о том, что человеку, опять же в отличие от компьютера, свойственно ошибаться.

Нередко приводят аргумент и другого рода. Он заключается в якобы имеющемся кризисе в области ИИ. Мы, однако, видели, что развитие в этой области шло путем последовательного повышения самостоятельности интеллектуальных систем. Самостоятельность алгоритмов поиска в пространстве решений распространялась лишь на конкретную задачу, в которой могли меняться только начальные данные. Методы представления знаний позволили компьютеру решать разные задачи, принадлежащие описанной человеком предметной области. Развитие машинного обучения позволило ослабить и эту зависимость от человека. Принципы воплощенного интеллекта подразумевают создание полностью автономных (хотя поначалу и сильно упрощенных) систем, которые обладают собственным поведением.

Кроме того, на стыке разных парадигм ИИ возникли и новые направления исследований: автоматическое формирование баз знаний в процессе обучения заставило исследователей всерьез заняться проблемой представления нечетких знаний, применение для которых методов поиска в пространстве решений (или методов манипулирования знаниями) рождает проблему рассуждения в условиях неопределенности. В рамках воплощенного интеллекта все прочие проблемы также приобретают дополнительное содержание: возникают вопросы обучения осмысленным понятиям, разработки когнитивных архитектур и т. д. «Периферийные» проблемы восприятия, речевого общения, управления, обычно рассматривающиеся весьма изолированно (им, к сожалению, и в этой книге не удалось уделить должного внимания, несмотря на их чрезвычайную важность), начинают все чаще изучаться в комплексе с «центральными» мыслительными процессами, и сейчас уже мало кто думает, что, скажем, систему распознавания произвольных изображений можно реализовать в качестве независимого модуля, который бы напрямую программировался и просто подключался к блоку мышления. Задачи, которые на сопоставимом с человеком уровне нельзя решить без построения сильного ИИ, иногда называют ИИ-полными. Перечень ИИ-полных задач, частичные решения которых традиционно создаются в рамках «слабого» искусственного интеллекта, все возрастает.

Ни одна из общих проблем в ИИ пока полностью не решена. Однако причина этого может быть понята именно в рамках алгоритмического подхода. От «сильного» ИИ компьютер отделяет трудность в реализации способности работать в алгоритмически полном пространстве: это относится и к проблеме поиска, и к проблеме представления, и к проблеме обучения (равно как и к проблемам восприятия, языка и управления). Конечно, речь не идет о том, чтобы ИИ мог гарантированно находить абсолютно произвольный алгоритм, решающий дедуктивную задачу или являющийся оптимальной моделью в индуктивной задаче: в любом случае на практике это невозможно хотя бы из-за ограниченности вычислительных ресурсов. Однако включать в рассмотрение лишь незначительную часть возможных решений даже низкой сложности, как это делается в методах «слабого» ИИ, для «сильного» ИИ недопустимо. Это объективная трудность, от которой нельзя спрятаться и в рамках «неалгоритмических» подходов. Игнорирование этой

трудности в проектах, ориентированных на создание сильного ИИ, приводит к тому, что цель этих проектов не может быть достигнута. Примером может служить «революционный» подход Джеффа Хокинса, описанный в его книге «Об интеллекте» в соавторстве с Сандрой Блейксли. Несмотря на то, что этот подход заимствовал некоторые мощные идеи (типа адаптивного резонанса) и имел заметную направленность на воспроизведение особенностей устройства человеческого мозга (преимущественно неокортекса), он не привел к существенному прогрессу, поскольку не преодолевал и даже не рассматривал главную трудность — работу в алгоритмически полном пространстве, без чего интеллект любой системы будет принципиально ограничен. Конечно, сейчас нельзя однозначно утверждать, что эта трудность в достаточной степени может быть разрешена в рамках самого алгоритмического подхода, но нет оснований предполагать и обратного, пока возможности алгоритмического подхода не исчерпаны. И в качестве одной из возможностей здесь видится исследование алгоритмических основ самоорганизации.

В физике нередко высказывалось мнение, что все наиболее интересное в ней уже открыто. Хоть это мнение всегда оказывалось неверным, оно продолжает периодически озвучиваться из-за того, что сделать действительно фундаментальное открытие здесь очень сложно. Аналогичное мнение в области ИИ сейчас вряд ли кто-то осмелится высказать из-за наличия сложных нерешенных проблем, которые делают работу в ней особенно интересной.

Стоит, правда, иметь в виду, что вряд ли в ИИ может быть найдена простая идея, которая решит все проблемы. Этим данная область отличается от физики, где фундаментальные законы имеют простой вид (низкую алгоритмическую сложность). Хотя можно надеяться, что некоторые наиболее фундаментальные принципы в ИИ тоже достаточно просты, их одних будет недостаточно для построения мыслящей машины (примерно в той же степени, в которой формулы $E = mc^2$ недостаточно для построения реактора, вырабатывающего электричество путем аннигиляции вещества). Это следует из того, что естественный интеллект обладает достаточно высокой алгоритмической сложностью и на его построение природой были затрачены огромные «вычислительные» ресурсы. Возможно, на высоком уровне абстракции разум — штука очень простая, но для реа-

лизации успешного поведения в реальном мире в наш мозг заложено множество специализированных механизмов (и не исключено, что вовсе не общие принципы, а именно эти механизмы и представляют собой то, что мы привыкли понимать под мышлением; в конце концов, сложность мышления обусловливается сложностью решаемых им задач, источником которых является внешний мир). В этой связи для создания ИИ придется либо затрачивать огромные вычислительные мощности для воспроизведения процесса самооптимизирующегося поиска, в результате которого возник естественный интеллект, либо вручную воплощать все необходимые механизмы (вероятнее всего, реализуется какой-то промежуточный вариант).

Логика глобальной эволюции подсказывает, что компьютеры (алгоритмические системы) стали ее неотъемлемой частью, очередным метасистемным переходом, а значит, они будут играть первоочередную роль в дальнейшем развитии, вероятнее всего, связанном с возникновением систем, находящихся на новом интеллектуальном уровне по сравнению с человеческим. Это является косвенным свидетельством возможности компьютерного воплощения разума. Конечно, имеются и альтернативы, например, новый интеллектуальный уровень может быть достигнут, если компьютер будет играть роль усилителя человеческого интеллекта (и этот сценарий в некотором смысле может быть предпочтительнее).

Уже давно компьютер стал неотъемлемой частью научных исследований, хотя и выполняет в них в основном низкоуровневые функции. Творческие функции остаются пока за человеком, ограниченные возможности которого, однако, становятся все более очевидными. Дальнейший прогресс науки подразумевает синтез многих знаний, но, как писал выдающийся ученый Ганс Селье в книге «От мечты к открытию», «...для того чтобы связать воедино многочисленные факты и прийти хоть к какому-то их пониманию, все они должны быть представлены в голове одного человека». Возможно, для человека это в принципе не по силам, и именно для этого может потребоваться ИИ (иногда говорится, что искусственным интеллектом занимаются те, кому не хватает своего естественного, с чем вполне можно согласиться, дополнительно отметив, что заметить нехватку интеллекта можно лишь используя его). Правда, и сама проблема создания ИИ носит аналогичный метасистемный характер.

Мы убедились в тесных взаимосвязях данной области с другими научными дисциплинами, к которым относятся не только когнитивные науки. И связи эти двусторонние. Если раньше высказывания о том, что «понятие сознания является одним из первичных и независимых понятий физического описания мира», можно было встретить в основном у философов (так писал М. Мамардашвили, обсуждая воззрения Декарта), то сейчас аналогичные утверждения встречаются и у физиков. В частности, со следующим высказыванием Р. Пенроуза во многом можно согласиться: «Научное мировоззрение, которое на глубинном уровне не желает иметь ничего общего с проблемой сознательного мышления, не может всерьез претендовать на абсолютную завершенность. Сознание является частью нашей Вселенной, а потому любая физическая теория, которая не отводит ему должного места, заведомо не способна дать истинное описание мира». Добавим лишь, что до сих пор все физические модели не выходили за рамки алгоритмического подхода.

Итак, исследования загадок мышления, которые сейчас трудно представить без попыток его алгоритмического описания и компьютерного воспроизведения, по своему влиянию на науку и технику, вероятно, окажутся наиболее значимыми. Иными словами, эти исследования заслуживают того, чтобы принять в них участие (правда, в скобках все же нужно отметить, что, по справедливому высказыванию Селье, не следует смешивать важность целей со значимостью самих исследований, которая зависит только от достигнутых результатов).

Абстрагирование *17, 116, 304, 435*
автомат
– клеточный *600–608, 616, 675–679*
– конечный *36, 183, 230, 231, 286, 525, 576*
агент интеллектуальный *109, 397, 509, 515, 523, 546, 646*
аксон *227–234, 252, 277, 279, 283*
алгоритм
– *k* внутригрупповых средних *331–335, 460*
– генетический *566–579, 586, 596, 598, 646*
– жадный *84, 89, 111, 420, 428, 437, 441, 584*
аллель *554, 586, 587*
амнезия *496, 499*
анимат *515–523, 527, 534, 538, 665*
анализ
– главных компонент *337–341*
– независимых компонент *344*
– факторный *57, 337–341*
архитектура когнитивная *513, 516*
астроцит *281, 282, 293*
аттрактор
– динамической системы *625–632, 640, 656–665*
– Лоренца *634, 635, 656*
– странный *276, 634–637, 641, 660*

Бифуркация *638–645, 691*
бритва Оккама *394, 395, 423*
бустинг *345*

Вероятность
– апостериорная (условная) *320, 324, 360, 444*
– априорная (безусловная) *321, 322, 325–327, 334, 352–387, 396, 423, 445*
взрыв комбинаторный *41–44, 102, 583, 584*
выборка
– контрольная (тестовая) *316, 350*
– обучающая *222–223, 272, 289, 299–344, 389, 390*
вывод
– дедуктивный *92, 120–157, 195, 218, 348, 362, 412, 422, 442–449, 507, 585*
– индуктивный *92, 155, 217, 348–353, 362, 368, 377–428, 442–449, 507, 585*
– немонотонный *138, 148–153*
вычисления
– мягкие *450*
– сверхтьюринговые *34, 287*
– символьные *101, 216, 240, 246*
– эволюционные *566–579*

Гипотеза
– лабиринтная *66–69, 75, 107, 119, 150, 483, 519*
– лингвистической относительности (Сепира—Уорфа) *158, 159, 219, 269, 328, 476, 685*
– физической символьной системы *28, 101, 211*
грамматика
– глубинная *184–189, 550*
– непосредственных составляющих *172–175, 182*
– стохастическая *184, 427, 443*
– трансформационная *185–187*
– формальная *175–184, 407, 421–442*

Дендрит *227–234, 283*
депривация *256, 270*
дилемма стабильности-пластичности *271, 361*

Жизнь искусственная *538–548, 576, 583, 584, 594, 600, 645, 648*

Задача
– NP-полная *40–52, 59, 76–89, 96–99, 104, 108–114, 132, 182, 246, 262, 273, 384, 420, 573–578, 584, 599*
– с неполными данными *332*
закон исключенного третьего *142*

Игра «Жизнь» *601–607, 647, 675, 678*
импринтинг *294, 301*
инвариант *242, 243, 246, 259, 273, 278, 302, 323, 338, 344, 481, 488, 506, 659*
интеллект искусственный
– воплощенный *507–514, 660, 702*
– слабый *9, 12, 32, 211, 704*
– сильный *9, 11, 28, 36, 52, 112, 119, 181, 211, 220, 225, 296, 513, 530, 531, 701–705*
интрон *555*
информации количество
– по Колмогорову *368–400*
– по Шеннону *342, 364–370, 374, 375, 610–611*

Квантор *19, 126–131, 144, 145, 455*
клетка
– ганглиозная *252–255*
– глиальная *228, 258, 280, 554*
– зрительной коры (простая, сложная) *259*
– пресинаптическая, постсинаптическая *228–234, 281*
когнитрон *259, 260, 281*
компьютер
– квантовый *43–52, 96, 595, 698*
– синергетический Хакена *659*
конкретизация *116, 193, 672*
коннективизм *240*
кубит *47–50*

Логика
– высказываний *120–125, 137, 144, 149, 154, 230, 286*
– интуиционистская *142, 143, 156*
– модальная *145, 194, 461, 464*
– нечеткая *449–462*
– предикатов *125–132, 194, 195, 672*
– предпочтений *147*
– темпоральная *146, 147*
– эпистемическая *145, 146*

Машина Тьюринга *23–52, 180, 181, 201, 226, 230, 285–293, 368, 378, 385, 399, 456, 607*
метаобучение *225, 576*
метаэвристика *578, 579, 585–595, 598, 600, 622, 649*
метод
– ближайшего соседа *301–306, 316, 335*
– имитации отжига *595–600, 646, 647, 651, 664*
– максимального правдоподобия *325, 326*
– Монте-Карло *91, 599*
– обратного распространения ошибки *241, 277, 282, 484, 599*
– ожидания-максимизации *332*
– опорных векторов *313–316*
– перекрестной проверки *316, 350*
– проб и ошибок *62*
– проблемных ящиков *62*
– эталонных образов *302–306, 308, 327, 332, 335*
множество Мандельброта *635*
модель
– N-грамм *165–175, 184, 192, 407–410, 429, 442*
– МакКаллока—Питтса *230, 235, 236, 275, 286*
морфогенез *603, 621*
мультиверс *698–700*

Набор правил *148–153, 208, 209, 412–421, 443*
негэнтропия *612–616, 619*
независимость статистическая *341, 367, 453, 454*
нейромедиатор *233, 278, 281, 534, 535*

Обучение
– активное *224, 349*
– без учителя *223, 257, 299, 328–336, 476, 488*
– инкрементное *225, 331, 442, 507, 545*
– латентное *484, 488, 590*
– с подкреплением *223, 483, 484, 522, 523*
– с учителем *223, 224, 299–316, 345*
– трансферное *400*
общий решатель задач *100–106, 119, 133, 207, 510*
онтогенез *217, 529, 537, 565*

Память
– ассоциативная *189, 243–250, 498, 659, 661*
– эйдетическая *497, 500*
парадокс
– Гемпеля *156, 445, 452*
– зелубых изумрудов *362, 387*
– лжеца *30, 140*
– лотерейный *444*
– Лошмидта *511*
параметр порядка *623, 631*
переменная
– лаговая *656*
– лингвистическая *457–461*
– пропозициональная *122*
переход метасистемный *681–689, 691–698, 706*
перцептрон *235–243, 244, 257, 277, 282, 286, 313, 383, 390, 599*
поведение
– адаптивное *515–523, 538, 539, 548, 565, 576*
– индуктивное *512*
– инстинктивное *60–63, 150, 492, 530, 549, 564*

подход
– аксиоматический *19–22, 35, 69*
– бионический *274*
– символьный *210*
– тестологический *55, 57, 340*
показатель Ляпунова *631, 637*
поле рецептивное *252–254, 258, 259, 283*
полнота алгоритмическая *24, 287, 390, 396, 398, 495, 649*
понятие довербальное *489, 490*
постоянная Фейгенбаума *644*
правило
– Байеса *321–326, 352, 355, 360, 380, 460, 465, 501*
– Локка *444, 445*
– решающее (классификационное) *299, 300, 305, 307, 316, 327*
– Хебба *235, 237, 245, 246, 258, 277, 281, 283, 484, 488, 498, 650, 664*
предположением о замкнутости мира *137–141*
премия Лёбнера *212*
принцип
– антропный *580–582, 587, 618, 697*
– минимальной длины описания *387–400, 402–405, 414–416, 420, 423, 460, 468, 506, 575, 652, 655*
проблема
– алгоритмически неразрешимая *29–38, 41–43, 46, 50–52, 93, 96, 104, 112, 132, 136, 180–182, 384, 547, 574, 702*
– останова *30–37, 51, 112, 180, 702*
программирование
– генетическое (эволюционное) *566, 569, 574–578, 588*
– объектно-ориентированное *199, 200, 202–206*
– процедурное *133, 134, 200, 202*
– функциональное *201, 688*
– эвристическое *76–83, 86–88, 90, 101, 106, 110–113, 117, 132, 208, 240, 246, 384, 415, 428, 510, 519, 542, 573, 578, 586*
проклятие размерности *38–52*
пропасть семантическая *222*
пространство
– признаков *298–310, 314–318, 322, 330, 334, 336, 337, 340–344, 392, 460, 471–477, 489, 507*
– поиска (вариантов, решений) *66, 78, 82, 102, 105, 110–114, 165, 509, 547, 566, 576, 578, 598, 704*
– фазовое *625, 626, 630–634, 641, 650, 652, 655, 657–667*
процедура
– альфа-бета отсечения (ветвей и границ) *88, 89, 100*
– минимакса *80, 85, 86, 92*
– направленного перебора *87–88, 599*
– порождающая *78–81*
– формирования рабочих оценок *85–89*
псевдослучайные числа *371, 373, 607, 668*
психология когнитивная *101, 119, 247, 296, 676, 703*

Равновесие Нэша *81, 542*
резонанс адаптивный *269–274, 362, 537, 697*
репеллер *625–629, 665*
речь
– аутическая *68, 470*
– внутренняя *68, 69*
– эгоцентрическая *68*
рефлекс *61–63, 78, 483–495, 498, 503, 522, 546, 548, 590, 682–686*
роботы
– адаптивные *508, 516, 664*
– интеллектуальные *508, 513*
– когнитивные *478, 513, 515, 523, 525, 538*
– с программным управлением *508*

Самообучение *224, 225, 257, 278–284, 594*
самоорганизация *236, 278, 279, 588, 595, 603, 604, 607, 618–623, 631, 645–650, 659, 664, 669, 674, 678–681, 703, 705*
семантика *195, 212, 216, 363, 474–482, 504, 513, 525, 530, 533, 537*
сеть
– ассоциативная *189–192*
– искусственная нейронная
– – Коско *247–249*
– – нейроглиальная *281–282, 289, 291, 293, 576*
– – прямого распространения сигнала *241, 244, 257, 484*
– – рекуррентная *244–249, 260, 269, 650, 652–654, 659, 661, 663*
– – спайковая *275*
– – Хопфилда *245–249, 659–661*
– семантическая *120, 192–197, 199, 218, 363, 474, 475, 478, 480*
силлогизм *122, 125, 128, 443*
синапс *228–234, 281–284*
сингулярность технологическая *688–700*
синергетика *279, 617–623, 659, 662, 666, 674*
синестезия *265*
система
– диссипативная *619–623, 631, 659, 662, 669*
– зрительная *8, 51, 56, 236, 250–262, 271, 283, 478, 500, 535, 680*
– мультиагентная *646–648*
– продукционная *148–153, 209, 416*
– экспертная *207–210, 216, 217, 220, 221, 240, 412, 442, 453, 459, 460, 467, 474, 508, 530, 649*
системогенез *679, 687*
сложность
– алгоритмическая *368–400, 455, 518, 575, 607–616, 633, 666*
– вычислительная *27, 38–43, 372, 384*
стратегия
– эволюционная *566–579*
– эволюционно устойчивая *542, 543, 592, 647*

Тавтология *123, 124, 131*
тезис Чёрча-Тьюринга *23, 28*
теорема Такенса *656–658*
теория
– Демпстера–Шафера *461–469*
– нечетких множеств *450–461, 577*
тест Тьюринга *211–215*

Уравнение Диофантово *20, 22, 31, 52*

Фитнесс-функция *566–579, 585, 586, 596*
фрактал *635, 636*
фреймы для представления знаний *196–207*
функция
– активационная *234, 238–240, 650, 653, 660*
– принадлежности *451–461*
– решающая *307–315, 328*

Хаос
– динамический (детерминированный) *633, 637–638, 641, 644–646, 664–666, 668*
– молекулярный *612–614*
химия супрамолекулярная *621*

Цикл предельный *629–631, 634, 637, 641–645, 660*

Эвристика
– поиска *72–76, 82, 83, 87, 90, 93, 99–114, 132, 151, 209, 384, 406, 425, 487, 523–526, 578, 587, 592, 649, 669, 680*
– разрешения конфликтов *151*
экзон *555*
эмерджентность *618, 621, 646, 647, 664, 672–681*
энтропия *342–344, 365, 366, 369, 381, 608–623*
эффект
– бабочки *638, 641*
– Болдуина *544*
– наблюдательной селекции *357*
– перцептивной готовности *268*
– чрезмерно близкой подгонки (переобучения) *309–316, 324–328, 334, 344–346, 352, 361, 377, 380, 383, 388, 390, 394, 414, 481, 494, 653*

H-теорема *611*
IQ-тест *55–59, 66, 103, 113, 320, 347, 560*

ISBN 978-5-7325-1008-9

НАУЧНОЕ ИЗДАНИЕ

Алексей Сергеевич **Потапов**

**ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
И УНИВЕРСАЛЬНОЕ МЫШЛЕНИЕ**

Главный редактор *Е. В. Шарова*
Редактор *Л. М. Манучарян*
Переплет художника *М. Л. Черненко*
Корректоры *Т. Н. Гринчук, А. А. Попова*
Компьютерная верстка *Т. М. Каргапольцевой*

Подписано в печать 15.05.2012. Формат издания 60×90 $^{1}/_{16}$.
Бумага офсетная. Гарнитура SchoolBookC. Печать офсетная. Усл. печ. л. 44,5.
Уч.-изд. л. 38,2. Тираж 2000 экз. Заказ

ОАО «Издательство «Политехника».
191023, Санкт-Петербург, Инженерная ул., д. 6.

Отпечатано с готовых диапозитивов
в ОАО «Издательско-полиграфическое предприятие
„Искусство России”.
191099, Санкт-Петербург, Промышленная ул., д. 38, корп. 2.