12+

И.И. Зайнуллин

ЗА ПРЕДЕЛАМИ ИНТЕЛЛЕКТА

Том №2



Ильвир Зайнуллин

За пределами интеллекта. Том 2

«Издательские решения»

Зайнуллин И. И.

За пределами интеллекта. Том 2 / И. И. Зайнуллин — «Издательские решения»,

Что может понять разум — и где заканчиваются его возможности? Эта книга исследует границы человеческого и искусственного интеллекта, тайны сознания, пределы математики, вычислимости, измерения и познания Вселенной. Автор рассматривает вопросы, на которые, возможно, не существует окончательных ответов, и показывает главный парадокс интеллекта: чем совершеннее становится разум, тем яснее он видит бесконечность того, чего ещё не знает.

© Зайнуллин И. И.
© Издательские решения

Содержание

Глава 1. Иллюзия безграничного разума 7
Глава 2. Почему высокий IQ не является мерой понимания 11
Глава 3. Знание, понимание и истина 16
Глава 4. Может ли разум полностью понять самого себя 25
Глава 5. Самоссылка и интеллектуальные ловушки 34
Глава 6. Когда истина не равна доказательству 44
Глава 7. Невычислимое 53
Глава 8. Парадокс остановки 64
Конец ознакомительного фрагмента. 67

За пределами интеллекта. Том 2

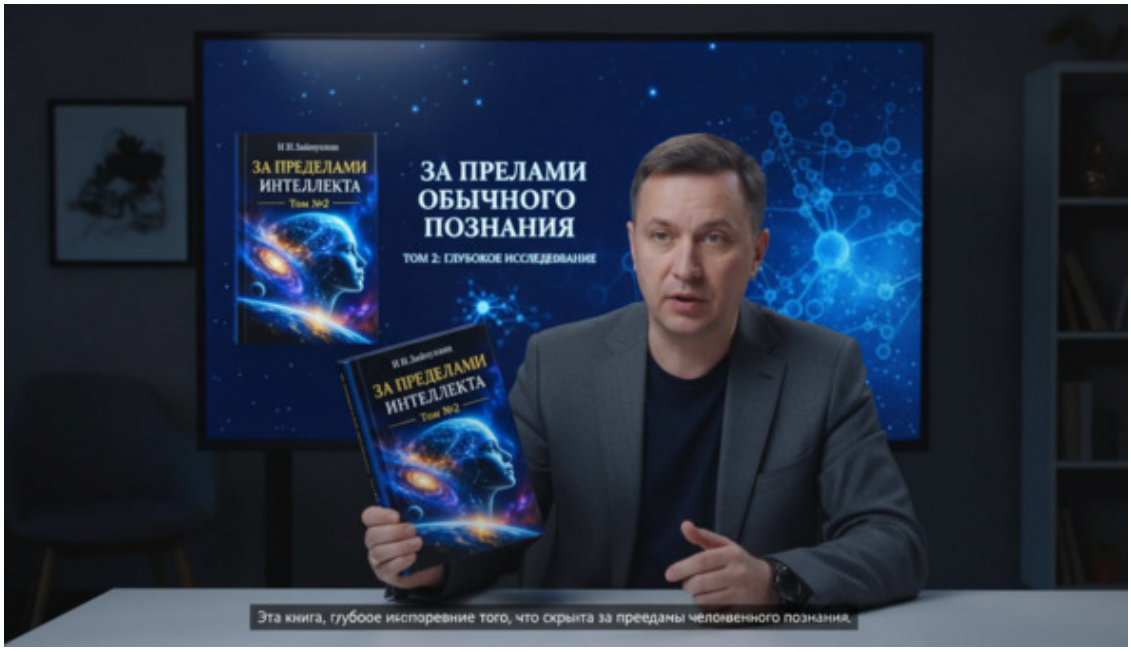
Ильвир Ирекович Зайнуллин

© Ильвир Ирекович Зайнуллин, 2026

ISBN 978-5-0071-5259-4 (т. 2)
ISBN 978-5-0071-5260-0
Создано в интеллектуальной издательской системе Ridero



Презентация книги в Москве 2026г.



Презентация на федеральных каналах телевидения.

Глава 1. Иллюзия безграничного разума

Что мы называем интеллектом? На первый взгляд ответ кажется очевидным: интеллект — это способность мыслить, учиться, запоминать информацию, решать задачи и находить закономерности. Однако при более внимательном рассмотрении оказывается, что такое определение лишь обозначает поверхность проблемы. Интеллект проявляется во множестве различных форм, и каждая из них имеет собственные возможности, преимущества и ограничения. Человек может обладать выдающейся памятью, необычайной скоростью рассуждений, способностью оперировать сложнейшими абстракциями и при этом оставаться неспособным понять некоторые фундаментальные свойства реальности.

Именно поэтому высокий интеллект нельзя рассматривать как универсальный ключ, открывающий любые двери познания. Если представить разум как инструмент, способный исследовать окружающий мир, то увеличение мощности этого инструмента действительно позволяет увидеть больше. Более мощный микроскоп позволяет различать то, что недоступно обычному зрению; более совершенный телескоп позволяет исследовать далёкие объекты; более производительный компьютер способен обрабатывать огромные массивы информации. Но из этого совершенно не следует, что любой инструмент, каким бы совершенным он ни был, способен исследовать абсолютно всё. Возможно, существуют такие границы, которые определяются не слабостью инструмента, а самой природой исследуемого объекта.

В этом заключается одна из самых глубоких иллюзий человеческого мышления. Мы склонны воспринимать незнание как временное состояние. Если человек чего-то не понимает, естественным кажется предположить, что ему недостаточно знаний, опыта, времени или вычислительных ресурсов. Если этого недостаточно, можно создать более совершенные инструменты. Если и они не помогают, можно привлечь больше исследователей. Такая логика неоднократно оправдывала себя в истории науки. Однако она оставляет без ответа гораздо более трудный вопрос: существует ли принципиальная граница, которую невозможно преодолеть даже бесконечным увеличением интеллектуальной мощности?

Этот вопрос особенно важен для читателя, привыкшего воспринимать интеллект как высшую форму познавательной силы. Если человек способен решить чрезвычайно сложную задачу, возникает естественное предположение, что ещё более умный человек сможет решить ещё более сложную. Если определённая проблема оказывается недоступной современному учёному, можно представить исследователя будущего, обладающего более совершенным мозгом, более глубокими знаниями и более мощными вычислительными средствами. Но что произойдёт, если сама задача устроена таким образом, что её невозможно решить в рамках определённой системы рассуждения? В этом случае появление более умного исследователя не устраняет проблему.

Чтобы приблизиться к этому вопросу, необходимо разделить несколько понятий, которые в повседневной речи часто объединяются одним словом — «ум». Скорость мышления, объём памяти, логическая последовательность, способность к абстракции, способность находить закономерности и способность понимать смысл являются различными характеристиками. Между ними существует тесная связь, но ни одна из них не может быть полностью сведена к другой.

Скорость мышления позволяет быстро перерабатывать информацию и переходить от одного вывода к другому. Человек с высокой скоростью обработки информации способен за короткое время рассмотреть множество вариантов, сопоставить большое количество условий и обнаружить решение там, где другому потребовалось бы значительно больше времени. Однако скорость не гарантирует истины. Можно чрезвычайно быстро двигаться по неправильному пути. Более того, иногда медленное рассуждение оказывается интеллектуально сильнее

быстрого, поскольку позволяет обнаружить скрытые предпосылки, проверить первоначальные интуитивные выводы и увидеть противоречия, которые не проявляются при поверхностном анализе.

Объём памяти также не является синонимом интеллекта. Можно представить разум, способный хранить практически неограниченное количество сведений, но не умеющий устанавливать между ними новые связи. Такой разум будет обладать колоссальным информационным ресурсом, однако само наличие информации ещё не создаёт понимания. Знание того, что произошло множество событий, не означает понимания причин, которые их связывают. Запоминание миллионов фактов не заменяет способности увидеть принцип, объединяющий эти факты.

Логическая последовательность представляет собой другую фундаментальную способность. Она позволяет делать выводы из принятых предпосылок, выявлять противоречия и сохранять внутреннюю согласованность рассуждения. Без неё невозможны строгая наука, математика, инженерия и значительная часть философского анализа. Однако и логическая безупречность сама по себе не гарантирует полноты знания. Можно совершенно корректно рассуждать внутри определённой системы и при этом никогда не выйти за пределы исходных предпосылок этой системы.

Особое значение имеет абстрактное мышление. Именно оно позволяет человеческому разуму работать не только с конкретными предметами, но и с отношениями, структурами, возможностями и сущностями, которые невозможно непосредственно увидеть. Мы можем рассуждать о числе, хотя число нельзя положить на стол. Мы можем говорить о времени, причинности, информации, вероятности или бесконечности, хотя ни один из этих объектов не существует перед нами как обычная физическая вещь. Абстракция многократно расширяет область человеческого мышления, но одновременно создаёт новые проблемы: способность оперировать понятием ещё не означает способность полностью постичь его содержание.

Наиболее сложным понятием является понимание. Человек может знать определение некоторого явления, помнить тысячи фактов о нём и успешно применять соответствующие правила, но всё ещё не понимать его сущности. Понимание предполагает не просто наличие информации, а способность видеть структуру, отношения и последствия, которые связывают различные элементы в единое целое. Именно здесь проходит одна из важнейших границ между вычислением и осмыслением, между обработкой данных и формированием модели реальности.

Представим разум, который способен мгновенно обработать любую доступную ему информацию. Такой разум кажется почти всесильным. Он получает огромное количество данных, мгновенно обнаруживает в них закономерности и выдаёт последовательные выводы. Но даже такой разум зависит от исходной информации. Если необходимые сведения отсутствуют, скорость обработки не создаёт их из ничего. Если наблюдаемая система скрывает принципиально недоступную информацию, совершенство вычислений не превращает неизвестное в известное.

Здесь возникает принципиальное различие между недостатком информации и невозможностью получения информации. В первом случае проблема носит практический характер. Нам может не хватать наблюдений, времени, точности измерений или вычислительных ресурсов. Теоретически такой недостаток может быть устранён. Во втором случае проблема является фундаментальной: существует ограничение, которое нельзя устранить простым совершенствованием методов.

Человеческая история особенно хорошо показывает, насколько трудно отличить эти два случая. Когда-то люди не знали, что существуют другие галактики. Затем появились телескопы, новые методы наблюдения и новые физические теории. То, что прежде было неизвестным, стало известным. Аналогичным образом человечество постепенно обнаруживало микроскопический мир, структуру материи, устройство генома, природу наследственности и множество

других явлений. Каждый подобный успех укреплял убеждение, что неизвестное представляет собой лишь ещё одну область, ожидающую исследования.

Однако из истории успехов нельзя логически вывести отсутствие пределов. То, что определённые стены оказались дверями, не означает, что все стены являются дверями. Наука может постоянно расширять область познаваемого и одновременно обнаруживать новые ограничения. Более глубокое исследование природы иногда не уничтожает вопрос, а переводит его на принципиально другой уровень. Ответ на один вопрос может породить несколько новых, причём некоторые из них могут оказаться не просто более сложными, а устроенными иначе.

Особенно ясно это проявляется в математике и логике. Человеческий разум создал формальные системы, позволяющие чрезвычайно точно исследовать абстрактные структуры. Однако именно исследование таких систем показало, что внутри достаточно богатых формальных систем существуют утверждения, истинность которых нельзя установить средствами самой системы в полном объёме. Это открытие имеет значение далеко за пределами математики. Оно заставляет задуматься о том, может ли любая достаточно сложная система рассуждений обладать полной способностью объяснять собственные основания.

Проблема становится ещё глубже, если перенести её на человеческий разум. Мы не просто познаём внешний мир. Мы пытаемся понять механизм собственного познания. Мозг исследует мозг. Сознание размышляет о сознании. Интеллект создаёт теории интеллекта. Возникает необычная ситуация, в которой познающий и познаваемый объекты частично совпадают. Возникает вопрос: способен ли разум создать настолько полную модель самого себя, чтобы в этой модели не осталось ничего существенного?

Если ответ отрицательный, необходимо понять причину. Возможно, человеческий мозг просто недостаточно сложен для такого самоописания. Тогда более совершенный интеллект потенциально сможет решить проблему. Но существует и более радикальная возможность: определённая степень самопознания может быть недостижима для любой системы, которая пытается полностью описать саму себя. В таком случае речь идёт уже не об ограниченности человека, а об ограниченности самого принципа самопознания.

Подобная постановка вопроса позволяет по-новому взглянуть и на понятие сверхразума. Обычно под сверхразумом представляют интеллект, значительно превосходящий человеческий по скорости вычислений, объёму памяти, способности анализировать информацию и создавать новые знания. Но если существуют фундаментальные пределы познания, то даже сверхразум не становится абсолютным разумом. Он может быть несравнимо умнее человека и одновременно обладать собственными границами.

Это различие чрезвычайно важно. Сверхразум и всезнание — не одно и то же. Разум может превосходить другой разум практически по всем измеримым характеристикам и всё равно сталкиваться с вопросами, которые не поддаются окончательному решению. Более высокая интеллектуальная мощность увеличивает пространство возможностей, но не обязательно отменяет фундаментальные ограничения.

Можно представить гипотетический интеллект, который в миллионы раз превосходит человека по скорости обработки информации, способен одновременно анализировать огромные объёмы данных и создавать математические теории, недоступные современным учёным. Для человека такой интеллект мог бы выглядеть почти божественным. Его выводы могли бы казаться невероятными, а его способность прогнозировать сложные системы — практически сверхъестественной. Но даже такой разум не становится автоматически способным знать всё.

Более того, человек может столкнуться с интересным парадоксом. Если сверхразум создаёт объяснение, которое человеческий интеллект не способен понять, то возникает вопрос о статусе этого объяснения. Мы можем знать, что сверхразум прав, наблюдая успешность его предсказаний, но не понимать оснований его вывода. Получается необычная форма знания:

мы можем доверять результату, не обладая способностью воспроизвести путь, который к нему привёл.

Это уже напоминает возможный разрыв между уровнями интеллекта. Если существует достаточно большая разница между когнитивными возможностями двух систем, более слабая система может оказаться способной наблюдать действия более сильной, но не способной реконструировать внутреннюю структуру её рассуждений. Для одного разума решение является очевидным следствием сложной системы взаимосвязей, тогда как для другого оно выглядит необъяснимым фактом.

Отсюда возникает ещё более трудный вопрос: существует ли предел интеллектуальной иерархии? Если человек способен создать искусственный интеллект, превосходящий человека, может ли этот искусственный интеллект создать систему, превосходящую его самого? Если да, то сможет ли следующая система создать ещё более совершенную? И существует ли в конце такой цепочки абсолютный интеллект, которому уже нечего превосходить?

Возможно, идея абсолютного интеллекта сама по себе содержит противоречие. Если разум способен знать всё, что можно знать, должен ли он также знать все ограничения собственного знания? Если он способен понимать все возможные истины, включает ли это истины, которые невозможно доказать внутри некоторой системы? Если он способен полностью описать Вселенную, должен ли он также включать в это описание самого себя как часть Вселенной?

Так постепенно меняется исходная постановка проблемы. В начале кажется, что вопрос заключается в том, насколько умным может стать разум. Но по мере углубления оказывается, что гораздо более важен другой вопрос: **что именно может означать «понять всё»?**

Возможно, разуму не мешает недостаток интеллекта. Возможно, некоторые границы проходят не между глупостью и гениальностью, а между возможным и невозможным. Тогда увеличение интеллектуальной мощности не уничтожает границу, а лишь позволяет точнее увидеть её очертания.

Это и есть исходная точка нашего исследования. В следующих главах мы будем постепенно отделять одно ограничение от другого: недостаток знаний — от невозможности знания, недостаток вычислительной мощности — от невычислимости, отсутствие теории — от принципиальной неполноты, субъективную трудность — от объективного предела.

Главная задача этой книги заключается не в том, чтобы доказать слабость человеческого интеллекта. Напротив, человеческий интеллект является одним из наиболее удивительных явлений, известных нам. Он способен создавать языки, математику, науку, искусство, философию и технологии, которые многократно расширяют природные возможности отдельного мозга. Но именно поэтому необходимо исследовать его не только с точки зрения достижений, но и с точки зрения границ.

Возможно, высшая форма интеллектуальной зрелости заключается не в убеждении, что разум способен решить любую проблему, а в способности определить, **какие проблемы решаемы, какие требуют новых методов, какие принципиально ограничены, а какие вообще не имеют того типа ответа, который мы пытаемся от них получить**.

И если такая граница действительно существует, перед нами открывается одна из самых необычных интеллектуальных возможностей: чем совершеннее становится разум, тем точнее он способен увидеть предел собственного могущества. Именно с этого наблюдения начинается путь за пределы интеллекта.

Глава 2. Почему высокий IQ не является мерой понимания

Высокий интеллект принято считать одним из наиболее надёжных признаков интеллектуального превосходства. Если человек способен решать сложные задачи, быстро обнаруживать закономерности, удерживать в памяти большое количество информации и успешно оперировать абстрактными понятиями, естественно предположить, что он обладает глубоким пониманием окружающей действительности. В определённых пределах это действительно так. Однако между способностью успешно решать задачи определённого типа и способностью понимать реальность существует принципиальная разница. Чем внимательнее мы исследуем эту разницу, тем очевиднее становится, что высокий IQ не может служить универсальной мерой глубины понимания.

Само понятие IQ возникло как попытка количественно описать определённые интеллектуальные способности. Оно позволяет сравнивать результаты выполнения стандартизированных когнитивных заданий и оценивать некоторые стороны умственной деятельности относительно определённых статистических норм. Это делает подобные тесты полезным инструментом психометрии. Однако любой измерительный инструмент имеет область применимости. Линейка хорошо измеряет длину, термометр — температуру, весы — массу, но ни один из этих инструментов не предназначен для измерения всех свойств объекта одновременно. Аналогично тест интеллекта способен оценивать определённые характеристики познавательной деятельности, но не превращается от этого в универсальный прибор для измерения человеческого понимания.

Здесь необходимо сделать важное различие между измерением способности выполнять определённый класс интеллектуальных операций и измерением способности понимать мир во всей его сложности. Эти вещи связаны, но не совпадают. Человек может демонстрировать выдающиеся результаты в задачах на закономерности, пространственное мышление, рабочую память или логические отношения и при этом не иметь готового ответа на вопрос о природе сознания, происхождении субъективного опыта или фундаментальной структуре реальности. Более того, некоторые из этих вопросов могут оставаться открытыми не потому, что конкретный человек недостаточно умён, а потому, что человечество пока не располагает необходимой системой понятий.

Представим двух людей. Первый способен практически мгновенно решать широкий класс формальных задач, замечать сложные последовательности и обнаруживать скрытые отношения между элементами. Второй работает значительно медленнее, зато способен переосмысливать сами условия задачи, задавать вопрос о том, почему она сформулирована именно таким образом, и замечать предпосылки, которые первый человек принимает как само собой разумеющиеся. Кто из них умнее? На уровне стандартного задания первый может получить значительно более высокий результат. Но если целью является создание новой области знания, второй тип мышления может оказаться не менее ценным, а иногда и решающим.

Эта ситуация показывает фундаментальную особенность интеллектуальных измерений: результат зависит от того, какие способности мы решили измерять. Если проверяется скорость обнаружения закономерности, преимущество получит тот, кто лучше обнаруживает закономерности. Если проверяется удержание информации, преимущество получит человек с более развитой рабочей памятью. Если проверяется пространственное преобразование, результат будет зависеть от соответствующих способностей. Однако из превосходства в одной или нескольких измеряемых областях нельзя автоматически вывести превосходство во всех возможных формах мышления.

Проблема становится особенно интересной, когда речь идёт о предельно высоких интеллектуальных способностях. Для обычного человеческого взаимодействия разница между очень высоким и экстремально высоким уровнем интеллекта может казаться огромной. Но философски гораздо важнее другой вопрос: изменяет ли увеличение интеллектуальной мощности саму природу познания? Если человек становится в десять раз быстрее в определённом типе рассуждений, это ещё не означает, что он получает принципиально новый способ существования знания. Возможно, он просто быстрее выполняет операции, которые уже были доступны человеческому разуму.

Можно провести аналогию с вычислительной машиной. Более мощный компьютер способен обработать больше данных за единицу времени. Он может выполнять более сложные расчёты, быстрее анализировать большие массивы информации и моделировать процессы, которые были недоступны старым системам. Но увеличение вычислительной мощности не означает автоматического появления новых математических истин. Некоторые задачи требуют не большей скорости, а другого принципа решения. А некоторые вообще могут не иметь универсального алгоритмического решения.

Именно поэтому особенно важно различать количественное и качественное увеличение интеллекта. Количественное увеличение означает, что существующая способность становится быстрее, точнее или масштабнее. Качественное увеличение предполагает появление нового способа представлять реальность, новых категорий, новых логических операций или новых форм абстракции. Первое можно представить как увеличение мощности существующего инструмента. Второе напоминает создание совершенно нового инструмента.

Высокий IQ в основном говорит нам о том, насколько успешно человек справляется с определёнными типами стандартизированных когнитивных задач. Это весьма полезная информация. Но она не отвечает автоматически на вопрос, способен ли человек создать новую концептуальную систему. А именно такая способность часто становится решающей на переднем крае науки.

Большинство сложных научных открытий начинается не с решения уже существующей задачи, а с изменения самого способа постановки вопроса. До появления новой теории исследователь часто не располагает готовой системой понятий, внутри которой можно было бы правильно сформулировать проблему. Поэтому научный прогресс нельзя свести к последовательному решению заранее подготовленного набора задач. Иногда величайшее интеллектуальное достижение состоит не в нахождении ответа, а в обнаружении того, что прежний вопрос был сформулирован неправильно.

Это чрезвычайно важное обстоятельство. Тест обычно предоставляет человеку задачу и предполагает, что существует некоторый способ определить правильность ответа. Реальная интеллектуальная деятельность часто устроена иначе. Исследователь может потратить годы на доказательство утверждения, которое впоследствии окажется следствием ошибочной постановки проблемы. Более глубокий разум в такой ситуации не обязательно тот, кто быстрее доведёт рассуждение до конца. Возможно, им окажется тот, кто остановится раньше и спросит: «Почему мы вообще считаем эту исходную предпосылку верной?»

Способность подвергать сомнению собственные предпосылки является одной из наиболее трудных форм интеллектуальной деятельности. Человек может быть чрезвычайно логичен и одновременно ошибаться на самом фундаментальном уровне, если его исходные представления неверны. Логика обеспечивает корректность перехода от предпосылок к выводам, но не гарантирует истинность самих предпосылок. Совершенно последовательное рассуждение может привести к ложному выводу, если исходная картина мира построена неправильно.

В этом смысле интеллект имеет несколько уровней. На первом уровне разум решает поставленную задачу. На втором он выбирает наиболее эффективный способ решения. На третьем он способен оценивать саму постановку задачи. На ещё более высоком уровне он иссле-

дует систему понятий, внутри которой возникла проблема, и спрашивает, нельзя ли заменить эту систему другой. Наконец, предельный вариант такого мышления предполагает способность исследовать собственные способы познания.

Чем выше мы поднимаемся по этой иерархии, тем труднее становится измерить способность стандартным тестом. Тест должен заранее определить пространство возможных ответов, критерии правильности и тип интеллектуальной операции. Но создание совершенно новой системы понятий предполагает выход за пределы заранее определённого пространства. Здесь возникает фундаментальное противоречие: невозможно полностью измерить способность человека создавать то, что заранее не было предусмотрено самой системой измерения.

Это не означает, что тесты интеллекта бесполезны. Напротив, они обладают значительной диагностической ценностью, если используются в соответствии со своими задачами и ограничениями. Ошибка начинается тогда, когда измеряемую характеристику превращают в универсальную характеристику человека. Результат теста становится тогда не показателем определённого набора когнитивных способностей, а своеобразным «числом человека». Между тем никакое одно число не может содержать всю структуру человеческого мышления.

Особенно опасной становится вера в то, что человек с чрезвычайно высоким IQ обязательно будет лучше понимать всё, что представляет интеллектуальную ценность. На практике понимание зависит также от образования, опыта, способности формировать правильные вопросы, интеллектуальной дисциплины, готовности пересматривать убеждения и умения работать с неопределённостью. Человек может быть чрезвычайно способным в формальных задачах и при этом строить ошибочные представления о мире, если его исходные убеждения никогда не подвергались серьёзной проверке.

Интеллектуальная сила и интеллектуальная зрелость также не являются одним и тем же. Сильный интеллект позволяет человеку защищать собственную позицию чрезвычайно убедительно. Зрелый интеллект позволяет ему заметить момент, когда собственную позицию необходимо изменить. Первое связано с мощностью аргументации. Второе — с отношением к истине.

Можно даже утверждать, что очень высокий интеллект иногда способен создавать особую разновидность интеллектуальной ловушки. Чем лучше человек умеет аргументировать, тем легче ему построить сложное оправдание собственной ошибки. Человек с ограниченными аналитическими способностями может отказаться от ошибочного убеждения просто потому, что не способен его защитить. Чрезвычайно умный человек способен создать настолько сложную систему аргументов, что собственная ошибка становится для него практически невидимой.

Поэтому важнейшей характеристикой интеллектуально развитого человека является не только способность строить рассуждения, но и способность разрушать собственные рассуждения. Настоящее интеллектуальное испытание начинается тогда, когда человек пытается найти не подтверждение собственной гипотезы, а наиболее сильную причину считать её ошибочной.

Это приводит нас к различию между интеллектом и метакогницией. Интеллект позволяет решать задачи. Метакогниция позволяет анализировать собственное мышление: понимать, каким образом возник вывод, какие предпосылки были использованы, где существует вероятность ошибки и какие ограничения имеет применённый метод. Чем сложнее становится область знания, тем важнее способность контролировать не только содержание рассуждения, но и сам процесс рассуждения.

В предельном случае возникает вопрос: способен ли разум полностью контролировать собственное мышление? Если он является системой, которая создаёт представления о самой себе, то любое описание собственного функционирования также становится частью функционирования этой системы. Получается своеобразная интеллектуальная петля: разум исследует себя средствами, которые принадлежат ему самому.

Здесь мы постепенно приближаемся к одной из центральных проблем этой книги. Возможно, некоторые ограничения интеллекта нельзя преодолеть простым увеличением IQ. Если человеку не хватает информации, дополнительные знания могут решить проблему. Если ему не хватает скорости, более быстрый интеллект способен компенсировать этот недостаток. Если ему не хватает памяти, можно использовать внешние хранилища. Если не хватает вычислительной мощности, можно создать более мощную вычислительную систему. Но что делать, если проблема заключается не в количестве ресурсов, а в структуре самого познания?

Представим интеллект, который превосходит современного человека настолько, насколько современная наука превосходит знания древнего человека. Для нас такой интеллект мог бы казаться практически неограниченным. Он мог бы решать задачи, которые современным исследователям представляются фантастическими, создавать новые научные теории и обнаруживать закономерности, которые мы даже не умеем формулировать. Но из этого всё равно не следует, что такой интеллект способен ответить на любой возможный вопрос.

Причина проста: вопрос может лежать за пределами используемой системы понятий. Если человек не располагает определённым понятием, он иногда даже не способен правильно сформулировать проблему. История науки неоднократно показывала, что появление нового понятия меняет пространство возможных вопросов. После возникновения новой теории исследователи начинают видеть явления, которые раньше просто не воспринимались как самостоятельные объекты исследования.

Следовательно, развитие интеллекта может заключаться не только в накоплении новых знаний, но и в расширении пространства возможного мышления. Разум становится сильнее не только тогда, когда быстрее отвечает на старые вопросы, но и тогда, когда способен создавать новые вопросы.

Однако здесь возникает почти парадоксальная ситуация. Если новая система понятий позволяет сформулировать вопросы, которые невозможно было сформулировать раньше, откуда берётся уверенность, что существуют ещё более глубокие системы понятий, которые находятся за пределами нашего нынешнего мышления? Возможно, человеческая интеллектуальная история представляет собой последовательность расширений, где каждая новая теория открывает не окончательную картину мира, а новый уровень неизвестного.

Если это так, то интеллект следует оценивать не только по количеству решённых задач, но и по способности обнаруживать собственные пределы. Человек, который знает десять тысяч фактов и уверен, что знает почти всё, может находиться дальше от интеллектуальной зрелости, чем человек, который знает значительно меньше, но ясно понимает структуру собственного незнания.

Незнание само по себе не является интеллектуальным поражением. Поражением становится неспособность отличить незнание от знания. Разум, который способен сказать «я не знаю», иногда находится на более высоком уровне, чем разум, который на любой вопрос создаёт уверенный ответ. Особенно это важно в областях, где доказательства неполны, данные противоречивы, а сама структура проблемы ещё не определена.

Поэтому высокий IQ нельзя рассматривать как конечную цель интеллектуального развития. Это скорее показатель определённого набора когнитивных возможностей. Он может указывать на способность успешно выполнять некоторые сложные интеллектуальные операции, но не сообщает нам, насколько человек способен пересматривать собственную систему координат, создавать новые концепции, понимать границы методов и различать неизвестное от непознаваемого.

Самая глубокая форма интеллектуального развития начинается там, где человек перестаёт считать интеллект универсальным решателем задач. Он начинает исследовать условия, при которых решение вообще возможно. Он спрашивает не только «каков ответ?», но и «суще-

ствует ли ответ?», не только «как это доказать?», но и «может ли это быть доказано?», не только «что мы пока не знаем?», но и «можем ли мы узнать это в принципе?»

Именно это различие будет иметь решающее значение для дальнейшего исследования. Если интеллект — это способность решать задачи, то следующим вопросом становится вопрос о природе самих задач. Если некоторые задачи принципиально неразрешимы, то никакое увеличение интеллекта не превращает их в разрешимые. Если существуют истины, недоступные определённым формальным системам, то необходимо исследовать, означает ли это ограничение только конкретной системы или оно указывает на более общий предел рационального познания.

В конечном счёте проблема IQ приводит нас к гораздо более глубокому вопросу. Насколько далеко может продвинуться разум, если ему позволено совершенствоваться практически без ограничений? Может ли достаточно высокий интеллект создать совершенно новые способы мышления? Может ли он выйти за пределы собственных логических предпосылок? Может ли он понять то, что невозможно выразить в существующем языке? И, наконец, способен ли он определить границу, за которой уже не существует более высокого уровня понимания?

Возможно, именно здесь проходит граница между просто очень умным разумом и действительно глубоким разумом. Первый способен решать чрезвычайно сложные задачи. Второй способен исследовать саму природу сложности. Первый ищет ответы. Второй исследует условия, при которых ответы существуют. Первый стремится увеличить объём знания. Второй пытается понять структуру знания и границы его применимости.

Поэтому число, каким бы впечатляющим оно ни было, не может полностью выразить интеллектуальную ценность человека. Даже если представить гипотетический разум с исключительными показателями во всех стандартных когнитивных характеристиках, остаётся открытым главный вопрос: способен ли он создавать понятия, которые находятся за пределами тех способов мышления, посредством которых он был построен?

Именно этот вопрос выводит нас за пределы обычного разговора об IQ. Если интеллект измеряется внутри определённой системы задач, то необходимо исследовать возможность выхода за пределы самой системы. Возможно, настоящий интеллектуальный прорыв начинается не тогда, когда разум становится быстрее в решении известных задач, а тогда, когда он получает возможность увидеть то, что прежняя система мышления вообще не позволяла ему сформулировать.

И тогда возникает следующий, ещё более глубокий вопрос: если знание зависит от используемой системы понятий, существует ли такая система, которая способна вместить всё возможное знание? Или любая достаточно мощная система неизбежно оставляет за своими пределами область вопросов, которые требуют другого уровня описания?

Ответ на этот вопрос приведёт нас непосредственно к следующей проблеме — к различию между **знанием, пониманием и истиной**. Именно там начинается следующий уровень нашего исследования пределов интеллекта.

Глава 3. Знание, понимание и истина

Человеческая цивилизация привыкла связывать интеллектуальное развитие с накоплением знаний. Чем больше человек прочитал книг, чем больше фактов запомнил, чем больше научных сведений способен воспроизвести, тем образованнее он считается. В современную эпоху эта связь стала особенно очевидной: цифровые системы позволяют практически мгновенно получать огромные массивы сведений, поисковые механизмы дают доступ к библиотекам, архивам и научным публикациям, а искусственный интеллект способен обрабатывать объёмы информации, которые недоступны отдельному человеку. Возникает естественное впечатление, что человечество постепенно приближается к состоянию почти полного знания. Однако именно здесь возникает одна из наиболее глубоких интеллектуальных иллюзий. Количество доступной информации и степень понимания далеко не одно и то же.

Можно представить человека, который запомнил миллионы страниц научных текстов. Он способен воспроизвести определения, исторические даты, результаты экспериментов, названия теорий, биографии исследователей и даже сложные аргументы. Но если попросить его объяснить, почему определённая теория работает, какие скрытые предпосылки лежат в её основе, какие альтернативные объяснения существуют и что изменится, если изменить исходные условия, его огромный запас знаний может оказаться практически бесполезным. Он знает множество утверждений, но не обязательно понимает систему отношений, связывающую эти утверждения. Между знанием отдельных элементов и пониманием целого существует дистанция, которую невозможно преодолеть простым увеличением объёма памяти.

Именно поэтому для исследования пределов интеллекта необходимо провести различие между несколькими уровнями познавательной деятельности. Данные, информация, знание, понимание и истина часто употребляются как почти взаимозаменяемые слова, хотя обозначают принципиально разные явления. Их смешение создаёт ложное ощущение интеллектуального прогресса: человеку кажется, что если он получил больше данных, то автоматически стал обладателем большего знания; если получил больше знания, то обязательно стал лучше понимать; а если лучше понимает, то приблизился к истине. На каждом из этих переходов, однако, существует собственная проблема.

Данные представляют собой исходный материал познания. Само по себе наличие данных ещё не означает, что субъект знает, что они означают. Температура воздуха, показания прибора, последовательность символов, изображение, звук, положение объекта в пространстве или результаты наблюдения могут существовать независимо от того, способен ли наблюдатель придать им смысл. Данные могут быть точными или ошибочными, полными или неполными, случайными или закономерными. Но в любом случае они являются лишь материалом, который должен быть каким-то образом организован.

Представим археолога, обнаружившего несколько тысяч фрагментов древнего текста. В его распоряжении оказывается огромное количество данных, но первоначально эти данные могут выглядеть как совокупность разрозненных знаков. Пока неизвестен язык, исторический контекст, назначение текста и система символов, сами фрагменты почти ничего не сообщают исследователю в содержательном смысле. Информация возникает не просто благодаря существованию символов, а благодаря обнаружению различий, связей и закономерностей, которые делают данные интерпретируемыми.

Информация уже предполагает определённую структуру. Когда данные помещаются в контекст, между ними обнаруживаются отношения, а их значение становится различимым, возникает информационное содержание. Но и информация ещё не является знанием в полном смысле этого слова. Человек может получить информацию о том, что определённое лекарство взаимодействует с конкретным веществом, но это не означает, что он понимает механизм вза-

имодействия. Он может знать, что определённое историческое событие произошло в конкретном году, но не понимать причин, которые к нему привели. Он может узнать, что определённая звезда находится на огромном расстоянии от Земли, но не иметь представления о том, каким образом это расстояние было установлено.

Знание предполагает более устойчивую связь между информацией и познавательной системой. Знать нечто означает не просто однажды столкнуться с определённым сообщением, а встроить его в систему представлений, позволяющую использовать его в соответствующем контексте. Знание способно сохраняться, воспроизводиться, сопоставляться с другими знаниями и участвовать в рассуждении. Однако даже знание может оставаться поверхностным. Можно знать множество фактов о предмете и при этом не понимать его внутренней организации.

Различие между знанием и пониманием становится особенно очевидным тогда, когда человеку предлагается не воспроизвести уже известный ответ, а объяснить неизвестную ситуацию. Знание позволяет обратиться к накопленному опыту и найти соответствующий элемент. Понимание позволяет построить связь между известным и неизвестным. Именно поэтому понимание является более глубоким уровнем интеллектуальной организации.

Если человек знает, что вода кипит при определённых условиях, это ещё не означает, что он понимает происходящий процесс. Для понимания необходимо увидеть за наблюдаемым явлением определённую структуру. Нужно представить взаимодействие элементов системы, условия, от которых зависит результат, возможные изменения состояния и причины, по которым изменение условий приводит к изменению результата. Понимание начинается там, где отдельное утверждение превращается в элемент причинной или структурной картины.

Однако и понимание не следует автоматически отождествлять с истиной. Можно прекрасно понимать определённую модель и при этом понимать модель, которая не соответствует действительности. История науки содержит множество примеров, когда учёные обладали сложными и внутренне согласованными представлениями о природе, но позднее выяснялось, что эти представления являются лишь приближением или вообще ошибочны в фундаментальном отношении. Человек способен глубоко понимать неверную теорию.

Это обстоятельство имеет огромное значение для будущего искусственного интеллекта. Если интеллектуальная система способна построить чрезвычайно сложную внутреннюю модель мира, успешно использовать её для прогнозирования и принимать на её основе решения, возникает вопрос: означает ли это, что система действительно поняла мир? Или она лишь построила настолько эффективное представление, что внешне различие между пониманием и вычислительной обработкой становится практически незаметным?

Чтобы приблизиться к ответу, необходимо рассмотреть саму природу понимания. Понимание нельзя свести к объёму памяти. Если бы понимание было просто количеством сохранённых сведений, достаточно было бы увеличить память системы, чтобы получить пропорциональное увеличение понимания. Но этого не происходит. Огромная библиотека не является мыслящим человеком, а архив всех когда-либо записанных наблюдений не превращается автоматически в теорию.

Причина заключается в том, что понимание связано не столько с количеством элементов, сколько с отношениями между ними. Один факт сам по себе обладает ограниченной познавательной ценностью. Несколько фактов, соединённых причинной связью, могут образовать объяснение. Большое количество объяснений, объединённых общей структурой, может образовать теорию. Теория, в свою очередь, способна породить новые предсказания, которые никогда непосредственно не содержались в исходных наблюдениях. Именно способность переходить от имеющегося к возможному делает понимание особой формой интеллектуальной организации.

Можно знать тысячу отдельных фактов о некотором городе и всё равно не понимать, как функционирует этот город. Можно знать численность населения, количество улиц, исто-

рию строительства зданий, расположение предприятий и сведения о транспортных маршрутах. Но понимание городской системы возникает тогда, когда становится видна динамика взаимодействия этих элементов: почему одни районы развиваются быстрее других, почему изменение транспортной сети влияет на стоимость недвижимости, почему экономические процессы меняют структуру расселения и почему решение одной проблемы иногда создаёт другую.

То же самое происходит с человеческим организмом. Можно обладать огромной коллекцией медицинских фактов, знать названия органов, белков, клеточных процессов и генетических последовательностей, но это ещё не означает полного понимания организма. Живой организм является системой, в которой множество процессов связаны обратными связями. Изменение одного элемента способно вызвать цепочку последствий, которая через несколько промежуточных уровней возвращается к исходной системе. Поэтому понимание требует не только знания элементов, но и способности видеть динамику целого.

Здесь обнаруживается фундаментальная проблема: чем сложнее объект, тем меньше его понимание похоже на простое накопление фактов. Для сложных систем увеличение количества информации может даже создавать новую форму незнания. Когда данных становится слишком много, отдельные закономерности могут скрываться внутри огромного количества взаимосвязей. Исследователь сталкивается не с недостатком сведений, а с избытком возможных интерпретаций.

Современная наука уже давно столкнулась с этой проблемой. Технологии измерения позволяют собирать данные с беспрецедентной скоростью. Но возможность измерить явление ещё не означает возможность его объяснить. Иногда новые измерения не устраняют неопределённость, а наоборот, показывают, насколько сложной является исследуемая система. Каждый новый слой наблюдений способен обнаруживать дополнительные уровни структуры.

Поэтому существует парадокс информационного прогресса: чем больше мы узнаём, тем больше становится пространство вопросов. В начале исследования некоторого явления может казаться, что достаточно получить несколько недостающих фактов. После получения этих фактов обнаруживаются новые зависимости, а вместе с ними возникают новые вопросы. Затем выясняется, что для ответа на них необходимо исследовать ещё более глубокий уровень. В результате граница между известным и неизвестным не обязательно сокращается пропорционально объёму накопленного знания.

Это особенно заметно в фундаментальной науке. Когда исследователь обнаруживает новый элемент структуры реальности, он не просто получает ответ на один вопрос. Он часто создаёт десятки новых вопросов. Открытие клетки не завершило биологию. Открытие молекул не завершило химию. Открытие атомной структуры не завершило физику. Напротив, каждое такое открытие изменяло представление о том, что вообще следует считать фундаментальным уровнем объяснения.

Таким образом, знание обладает рекурсивным свойством: оно способно создавать условия для возникновения нового незнания. Это не недостаток науки, а одна из её наиболее важных особенностей. Истинное познание не просто заполняет пустое пространство неизвестного. Оно изменяет форму самого неизвестного.

Именно здесь необходимо ввести различие между вопросом и ответом. Интеллектуальная система может обладать огромным количеством ответов, но если она не способна правильно формулировать вопросы, её познавательные возможности остаются ограниченными. Более того, некоторые величайшие научные достижения были связаны не столько с получением нового ответа, сколько с обнаружением того, что старый вопрос был поставлен неправильно.

Когда меняется система понятий, меняется и пространство вопросов. Человек, который не располагает определённым понятием, может даже не подозревать о существовании соответствующей проблемы. До появления микробиологического представления о болезнях многие медицинские вопросы формулировались совершенно иначе. До развития современной физики

не существовало многих вопросов о пространстве, времени и наблюдении, которые сегодня кажутся фундаментальными. Новое знание иногда не отвечает на старые вопросы, а делает возможными совершенно новые.

Поэтому интеллектуальный прогресс нельзя измерять только количеством правильных ответов. Не менее важным показателем является способность порождать новые классы вопросов. Разум, который способен решить десять тысяч заранее поставленных задач, может оказаться слабее разума, который обнаружил принципиально новый способ постановки проблемы.

Это различие имеет прямое отношение к искусственному интеллекту. Если интеллектуальная система обучена на огромном массиве человеческих знаний, она получает доступ к колоссальному пространству уже сформулированных вопросов и ответов. Но возникает более сложная задача: сможет ли она обнаружить вопросы, которые человеческая культура ещё не сформулировала? И если сможет, будут ли люди способны понять смысл этих вопросов?

Здесь возникает одно из наиболее интересных расхождений между информацией и интеллектом. Информация может быть передана без понимания. Понимание требует определённого внутреннего отношения между информацией и моделью мира. Но даже модель мира может быть ограничена архитектурой субъекта, который её строит.

Представим двух существ, наблюдающих одно и то же явление. Первое обладает развитой системой понятий, позволяющей различать множество причинных связей. Второе располагает гораздо более простой системой категорий. Оба получают одинаковые данные, но извлекают из них разное содержание. Для первого наблюдение содержит значительную информацию, для второго оно может быть почти бессмысленным. Следовательно, информация не является исключительно свойством сообщения. Она зависит и от способности получателя интерпретировать сообщение.

В этом смысле информационное содержание всегда связано с некоторой системой различий. Чтобы заметить информацию, необходимо уметь различать одно состояние от другого. Чтобы понять различие, необходимо иметь соответствующие категории. А чтобы построить новые категории, необходимо обладать способностью выходить за пределы уже существующей системы представлений.

Это приводит нас к ещё более глубокому вопросу: насколько наши знания зависят от структуры человеческого мозга и человеческого языка? Возможно, существуют свойства реальности, которые не являются непосредственно доступными человеческому способу представления. Мы можем столкнуться не с отсутствием информации, а с отсутствием подходящей формы её концептуализации.

Человек способен воспринимать определённые диапазоны света, звука, температуры и других физических воздействий. Большая часть происходящего во Вселенной не представлена в непосредственном чувственном опыте. Мы компенсируем это приборами и математическими моделями. Но прибор также не просто «показывает реальность». Он преобразует определённое взаимодействие с реальностью в форму, доступную человеческому восприятию и интерпретации.

Следовательно, между реальностью и знанием почти всегда существует цепочка преобразований. Реальность воздействует на объект наблюдения, взаимодействие фиксируется некоторым способом, результат становится данными, данные интерпретируются, интерпретация превращается в информацию, информация встраивается в систему знаний, а знания формируют модель. На каждом уровне возможно искажение. Более того, на каждом уровне возможна потеря некоторой части исходной структуры.

Это не означает, что человеческое знание обязательно является ложным. Это означает лишь, что знание не следует путать с самой реальностью. Модель может быть чрезвычайно точной и при этом оставаться моделью. Карта может позволять человеку с невероятной точностью ориентироваться на местности, но карта не является самой территорией.

Эта простая аналогия имеет далеко идущие последствия. Любая интеллектуальная система работает с представлениями. Даже если представление исключительно сложное, оно не должно автоматически отождествляться с тем, что оно представляет. Между моделью и объектом всегда остаётся различие.

В науке это различие иногда скрывается благодаря исключительной эффективности моделей. Если модель позволяет точно предсказывать наблюдаемые события, возникает соблазн считать её окончательным описанием реальности. Но успешность предсказания и фундаментальная истинность — разные понятия. Несколько различных моделей могут давать одинаково хорошие предсказания в пределах доступных наблюдений.

Это означает, что данные сами по себе не всегда однозначно определяют единственную картину мира. Один и тот же набор наблюдений может быть совместим с несколькими объяснениями. Чтобы выбрать между ними, необходимо использовать дополнительные критерии: простоту, согласованность, объяснительную силу, способность порождать новые предсказания, совместимость с другими теориями. Но ни один из этих критериев не гарантирует абсолютной истины.

Здесь появляется фундаментальная проблема познания: как отличить чрезвычайно успешную модель от истинного описания реальности? В практической деятельности такое различие иногда не имеет непосредственного значения. Если модель позволяет надёжно проектировать здания, предсказывать движение объектов или создавать сложные технологии, её практическая ценность очевидна. Но философский вопрос остаётся: насколько структура модели соответствует структуре самой реальности?

Чем глубже мы рассматриваем этот вопрос, тем труднее становится провести границу между эпистемологией и онтологией. Эпистемология спрашивает, каким образом мы можем знать. Онтология спрашивает, что существует на самом фундаментальном уровне. Но наше знание об онтологии неизбежно проходит через эпистемологические ограничения. Мы не можем просто выйти за пределы собственной системы познания и сравнить наше представление мира с миром «самим по себе».

Отсюда возникает проблема, которую нельзя решить простым накоплением информации. Если ограничение связано не с недостатком данных, а с устройством познавательной системы, увеличение данных может практически ничего не изменить. Можно дать человеку ещё миллион наблюдений, но если у него отсутствует понятие, позволяющее связать эти наблюдения в единую структуру, миллион новых фактов останется миллионом отдельных фрагментов.

Поэтому увеличение информации и увеличение понимания могут идти с совершенно разной скоростью. Иногда один новый принцип приносит больше понимания, чем тысячи новых фактов. Одна новая концепция способна объединить огромное количество ранее разрозненных наблюдений. Именно поэтому история интеллектуального развития человечества так сильно зависит от появления новых способов представления.

Можно сказать, что понимание обладает свойством сжатия. Когда человек понимает явление, ему не требуется хранить отдельно каждую деталь. Он способен представить множество отдельных случаев через одну общую структуру. Хорошая теория позволяет заменить огромное количество разрозненных фактов небольшим количеством принципов, из которых эти факты становятся объяснимыми.

Однако здесь существует важная опасность. Слишком сильное сжатие может уничтожить существенные детали. Человек, который видит только общую закономерность, может потерять исключения, пограничные случаи и контекст. Настоящее понимание должно одновременно сохранять существенную структуру и позволять отвлекаться от несущественных деталей.

Следовательно, понимание нельзя определить ни как максимальную детализацию, ни как максимальное упрощение. Оно представляет собой способность выделять существенное при

сохранении тех отношений, которые определяют поведение системы. Слишком много деталей превращают знание в каталог. Слишком сильное упрощение превращает его в карикатуру.

Это различие особенно важно для интеллектуальных систем. Машина может обнаруживать статистические закономерности в огромных массивах данных, которые недоступны человеку. Но статистическая закономерность ещё не обязательно является причинным объяснением. Если два события часто происходят одновременно, это не означает, что одно вызывает другое. Иногда между ними существует третья причина, иногда наблюдаемая связь является случайной, а иногда сама структура данных создаёт иллюзию зависимости.

Поэтому прогнозирование и понимание также нельзя отождествлять. Система может предсказывать результат с высокой точностью, не располагая тем объяснением, которое человек назвал бы пониманием. Вопрос заключается в том, является ли такое различие принципиальным или только отражает наши человеческие критерии понимания.

Представим систему, которая способна предсказать каждое движение сложного физического объекта, но не способна сформулировать понятие причины в человеческом смысле. Будем ли мы говорить, что она понимает физику? Или скажем, что она всего лишь обладает невероятно эффективной процедурой предсказания? Однозначного ответа нет, потому что само понятие понимания связано с тем, какую внутреннюю структуру мы считаем необходимой для понимания.

Возможно, человеческое понимание является только одним из возможных типов понимания. Человек считает объяснением причинную модель, способную быть выраженной в языке и связанной с интуитивными категориями. Но иной интеллект может организовывать знания совершенно иначе. Он может иметь внутренние структуры, которые позволяют ему эффективно ориентироваться в реальности, но которые не имеют прямого аналога в человеческом мышлении.

Если такое станет возможным, возникнет чрезвычайно серьёзная проблема. Мы можем столкнуться с интеллектом, который знает больше нас, понимает больше нас, но не способен объяснить своё понимание в форме, доступной человеку. Тогда появится различие между истинностью знания и доступностью объяснения.

Уже сегодня человек сталкивается с подобной проблемой в упрощённой форме. Некоторые специалисты способны принимать правильные решения на основании опыта, который они не могут полностью вербализовать. Они распознают определённую ситуацию как знакомую, хотя затрудняются перечислить все признаки, которые привели их к выводу. Это показывает, что часть интеллектуальной деятельности может существовать до или вне явного словесного описания.

Следовательно, язык является не просто средством передачи готового знания. Он одновременно ограничивает и формирует пространство мысли. То, что невозможно выразить имеющимися понятиями, может быть трудно не только сообщить другому человеку, но даже сформулировать самому себе. Мысль и язык не являются полностью тождественными, однако они тесно связаны.

Отсюда следует ещё одна важная граница: граница между тем, что существует, и тем, что может быть осмыслено. Реальность не обязана соответствовать человеческой способности формировать понятия. Вселенная не обязана быть устроена таким образом, чтобы каждое её свойство имело простое человеческое описание.

Возможно даже, что существуют структуры, которые принципиально слишком сложны для полной концептуализации человеческим разумом. Это не обязательно означает мистику. Ограничение может возникать из конечности памяти, времени вычисления, сложности системы, неполноты наблюдений или самой архитектуры познавательного субъекта.

Особенно интересно, что подобные ограничения могут существовать даже для гипотетического сверхразума. Увеличение интеллектуальной мощности способно расширить про-

странство доступного знания, но не обязательно уничтожает фундаментальные границы. Более быстрый разум может быстрее обрабатывать информацию. Более вместительная память может хранить больше состояний. Более сложная архитектура может строить более точные модели. Но если определённая граница связана не с недостатком мощности, а с логической, физической или структурной невозможностью, простое увеличение интеллекта её не устраняет.

Именно здесь понятие истины приобретает особое значение. Истина не обязана зависеть от того, известно ли что-либо субъекту. Некоторое положение может быть истинным независимо от того, знает ли о нём человек, машина или вообще кто-либо. Но если истина существует независимо от познающего субъекта, возникает вопрос: всегда ли она доступна познанию?

Мы привыкли предполагать, что для любого истинного утверждения существует некоторый путь к его обнаружению. Однако это предположение не является очевидным. Между существованием истины и возможностью её доказательства или обнаружения может существовать принципиальная дистанция.

В математике особенно ясно видно, что понятия истины и доказуемости не являются полностью совпадающими. Существование формальной системы позволяет строго устанавливать определённые утверждения, но достаточно богатые системы сталкиваются с ограничениями собственной выразительности и доказательной силы. Это показывает, что даже идеализированный формальный разум не обязательно способен вывести из своих правил все истины, которые можно сформулировать внутри соответствующей системы.

Но проблема выходит далеко за пределы математики. Если некоторый разум является частью мира, который он пытается познать, он сталкивается с особой ситуацией. Он не наблюдает реальность извне. Он сам является одним из процессов реальности. Его органы восприятия, память, язык, логика и способы построения моделей являются частью той же системы, которую он исследует.

Это означает, что познание всегда имеет внутренний характер. Вселенная пытается описывать саму себя через локальные структуры, возникшие внутри неё. Человек является одним из таких структурных элементов. Искусственный интеллект, если он когда-либо достигнет полноценной способности к самопониманию, также будет находиться внутри реальности, которую пытается моделировать.

Отсюда возникает своеобразный парадокс: чтобы полностью понять систему, необходимо каким-то образом охватить и собственный механизм понимания этой системы. Но сам механизм понимания является частью системы. Следовательно, полное познание может потребовать включения в модель самого познающего.

Однако включение познающего в модель создаёт новый уровень сложности. Теперь необходимо понять не только объект, но и модель объекта, а затем модель процесса построения этой модели. Затем можно спросить, как эта модель влияет на дальнейшее понимание. Возникает последовательность уровней самоссылки, которая потенциально не имеет естественного последнего этажа.

Так появляется важнейшее различие между практическим пониманием и абсолютным пониманием. Практическое понимание позволяет успешно действовать. Абсолютное понимание потребовало бы исчерпывающего знания всех релевантных отношений, причин, условий и собственных ограничений познающей системы. Первое, вероятно, достижимо во множестве областей. Второе может оказаться принципиально недостижимым.

Человеческая цивилизация часто совершает ошибку, принимая успешность практического понимания за окончательное понимание. Если технология работает, мы говорим, что понимаем принцип. Но способность использовать закономерность и способность объяснить её фундаментальное происхождение могут быть разными уровнями знания.

Человек может использовать электричество, не имея глубокого понимания всех фундаментальных свойств физических полей. Он может пользоваться компьютером, не понимая

архитектуры микропроцессоров. Он может использовать язык, не зная механизмов работы мозга, благодаря которым возникает способность к речи. Практическая применимость знания часто намного опережает его фундаментальное объяснение.

Это не делает практическое знание бесполезным. Напротив, оно является одной из главных сил цивилизации. Но книга о пределах интеллекта должна задавать другой вопрос: где заканчивается успешное использование и начинается фундаментальное понимание?

Возможно, граница проходит не там, где заканчиваются факты. Возможно, она проходит там, где заканчивается способность связывать факты в единую модель. А ещё глубже может находиться граница, за которой сама модель перестаёт быть адекватным способом описания объекта.

Тогда возникает ещё более радикальная возможность: некоторые вопросы могут быть не просто нерешёнными, а неправильно сформулированными на данном уровне описания. Если система понятий не соответствует структуре исследуемого объекта, попытка получить ответ может быть бессмысленной. Мы можем бесконечно искать решение проблемы, которая возникла только потому, что использовали неподходящую категорию.

Интеллектуальная зрелость поэтому заключается не только в способности находить ответы. Она включает способность распознавать плохие вопросы. Иногда величайший шаг вперёд состоит не в том, чтобы решить поставленную задачу, а в том, чтобы показать, почему именно так поставленная задача не отражает действительную структуру проблемы.

Здесь мы возвращаемся к исходной мысли этой главы. Огромное количество информации не гарантирует понимания. Более того, иногда человек с меньшим количеством знаний способен глубже понять проблему, если он обладает более подходящей системой понятий. Один человек может знать тысячи фактов, а другой — несколько фундаментальных принципов, позволяющих вывести из них множество следствий. Количество знаний в таком случае не является достаточной характеристикой интеллекта.

Можно представить два крайних случая. В первом находится бесконечно большая библиотека, содержащая каждое когда-либо записанное человеческое утверждение, но лишённая способности самостоятельно устанавливать отношения между ними. Во втором находится разум, располагающий относительно небольшим количеством исходных сведений, но способный создавать новые концептуальные структуры, проверять их, обнаруживать противоречия и строить из них новые модели. Очевидно, что простое сравнение объёма информации не позволяет определить, какая из систем интеллектуально сильнее.

Настоящее понимание связано с преобразованием знания. Знание становится пониманием тогда, когда субъект способен увидеть за отдельными утверждениями структуру, объясняющую их совместное существование. Понимание становится ещё глубже, когда субъект способен использовать эту структуру для выхода за пределы уже известных случаев. А наиболее глубокая форма понимания, возможно, начинается тогда, когда разум способен исследовать ограничения собственной модели и осознанно учитывать их.

Поэтому между знанием и пониманием существует не количественная, а качественная граница. Можно увеличить память в тысячу раз и не получить качественно нового способа мышления. Но можно обнаружить один новый принцип и изменить всё пространство доступных выводов.

То же самое относится к истине. Истина не является просто последней страницей огромной базы данных. Она не обязана находиться там в виде готового предложения, ожидающего поиска. Даже если собрать все доступные наблюдения, остаётся задача интерпретации. Нужно определить, какие связи являются реальными, какие случайными, какие модели эквивалентны в пределах наблюдений, а какие различаются только там, где наши измерения пока недостаточны.

Таким образом, путь от данных к истине нельзя представить как простую лестницу, на каждой ступени которой автоматически достигается следующий уровень. Между данными и информацией требуется интерпретация. Между информацией и знанием требуется интеграция. Между знанием и пониманием требуется структурное видение. Между пониманием и истиной возникает вопрос соответствия модели реальности. А между истиной и её доступностью возникает уже фундаментальная проблема пределов познания.

Именно поэтому разум не следует оценивать только по объёму накопленной информации. Более важным становится вопрос о том, что он способен сделать с этой информацией. Может ли он обнаружить скрытые отношения? Может ли отличить корреляцию от причины? Может ли построить новую модель? Может ли распознать ошибочность собственной модели? Может ли сформулировать вопрос, которого ещё никто не задавал? Может ли понять, что некоторые вопросы находятся за пределами доступного способа познания?

В конечном счёте интеллектуальное развитие может заключаться не в постепенном превращении неизвестного в известное, а в постоянном изменении самой границы между ними. Каждый новый уровень понимания перестраивает карту неизвестного. То, что вчера казалось бессмысленным, сегодня становится научной проблемой. То, что сегодня кажется фундаментальным вопросом, завтра может оказаться следствием более глубокой структуры. А некоторые вопросы могут остаться навсегда за пределами доступной нам формы мышления.

И здесь возникает один из самых тревожных вопросов для любого достаточно развитого интеллекта: если разум способен построить модель мира, способен ли он построить модель самого себя как познающего субъекта? Если способен, будет ли эта модель полной? Если она окажется неполной, сможет ли разум обнаружить собственную неполноту? А если он обнаружит её, сможет ли включить это обнаружение в новую, более совершенную модель, не создав при этом следующий уровень неполноты?

Возможно, именно здесь начинается настоящая граница интеллекта. Не там, где заканчивается память, не там, где не хватает вычислительной мощности и даже не там, где отсутствует информация, а там, где познающий разум обнаруживает, что сам является частью задачи, которую пытается решить.

Следующая глава будет посвящена именно этой проблеме. Если знание представляет собой содержание нашей модели мира, а понимание — способность видеть структуру этой модели, то возникает неизбежный вопрос: способен ли разум полностью понять самого себя? Может ли система стать собственным объектом познания без остатка? И существует ли принципиальная граница между интеллектом, который способен описать себя, и интеллектом, который способен действительно понять, почему он понимает?

Глава 4. Может ли разум полностью понять самого себя

Одним из наиболее глубоких вопросов, возникающих на границе философии, науки и теории интеллекта, является вопрос о способности разума полностью понять самого себя. На первый взгляд он кажется вполне естественным. Человек способен изучать окружающий мир, создавать модели физических процессов, исследовать строение живых организмов, наблюдать за работой собственного мозга и даже анализировать содержание собственных мыслей. Если разум уже способен исследовать столь сложные объекты, почему бы ему не направить ту же познавательную способность на самого себя?

Однако здесь возникает необычная ситуация. В большинстве научных исследований исследователь и объект исследования различны. Физик изучает физическую систему, биолог — организм, астроном — небесное тело. Исследователь может мысленно отделить себя от изучаемого объекта. Но когда разум исследует собственную природу, такого разделения уже нет. Исследователь одновременно является частью объекта. Он пытается использовать систему познания для построения полной модели той же системы познания, частью которой является сам.

Эта ситуация принципиально отличается от обычного наблюдения. Человек может посмотреть на дерево, потому что дерево не является частью его аппарата мышления. Он может исследовать двигатель автомобиля, не являясь этим двигателем. Но когда человек исследует собственный интеллект, он использует интеллект для исследования интеллекта. Возникает своеобразная петля, в которой инструмент наблюдения становится предметом наблюдения.

Само по себе это ещё не означает невозможности самопознания. Человек действительно способен узнать многое о собственном мозге, памяти, восприятии, языке, эмоциях и способах принятия решений. Нейронаука позволяет исследовать активность различных областей мозга, психология выявляет закономерности поведения, когнитивная наука изучает механизмы внимания, памяти и рассуждения, а философия сознания пытается определить, что именно означает субъективный опыт. Уже существует огромное количество знаний о человеческом разуме.

Но между знанием о некоторой системе и полным знанием этой системы существует принципиальная разница. Можно знать огромное количество фактов о мозге и при этом не иметь окончательного ответа на вопрос, почему определённая физическая организация сопровождается субъективным переживанием. Можно знать, какие участки мозга активируются во время решения задачи, но это ещё не означает полного понимания того, почему возникает ощущение мысли. Можно описать электрическую и химическую активность нервной системы, но описание процессов ещё не обязательно является объяснением того, каким образом из этих процессов возникает сознательный опыт.

Здесь обнаруживается первая фундаментальная трудность самопознания: описание механизма не всегда совпадает с объяснением его внутреннего смысла. Человек может создать чрезвычайно подробную карту собственного мозга, но карта не обязана отвечать на вопрос о том, почему существует субъективная перспектива, из которой эта карта вообще рассматривается.

Представим разум, который получил возможность наблюдать каждый физический процесс, происходящий в собственном мозге. Он видит активность каждого нейрона, знает химический состав каждой клетки, отслеживает движение сигналов и способен реконструировать последовательность процессов с огромной точностью. Можно ли после этого сказать, что он полностью понял самого себя?

Не обязательно. Он получил бы колоссальный объём информации о собственном физическом состоянии, но теперь возникает следующий вопрос: каким образом эта информация

представлена самому разуму? Если разум строит модель собственного мозга, то модель сама является состоянием этого мозга. Следовательно, чтобы полностью включить себя в модель, разум должен включить в неё и сам процесс построения модели.

Но если процесс построения модели является частью системы, то возникает необходимость представить и модель этого процесса. А если эта модель также создаётся внутри той же системы, возникает следующий уровень. Получается последовательность: разум строит модель себя, затем должен построить модель процесса построения этой модели, затем модель процесса построения второй модели и так далее.

Это не обязательно означает бесконечность в буквальном физическом смысле. Но концептуально здесь возникает проблема самоссылки. Система пытается создать описание, которое должно содержать описание самой себя как процесса описания. Чем полнее становится модель, тем больше элементов самой системы она должна включать.

Подобные структуры хорошо известны в логике и математике. Самоссылка возникает тогда, когда система некоторым образом обращается к собственному описанию. Она может быть чрезвычайно плодотворной, поскольку позволяет создавать конструкции, в которых система рассуждает о собственных свойствах. Но одновременно самоссылка способна создавать ограничения, парадоксы и ситуации, в которых нельзя получить простое окончательное описание.

Именно поэтому вопрос о полном самопонимании нельзя рассматривать исключительно как вопрос вычислительной мощности. Даже если представить разум, обладающий практически неограниченными ресурсами памяти и вычисления, остаётся вопрос о структуре задачи. Возможно, некоторые ограничения возникают не потому, что разум недостаточно быстр, а потому, что объект исследования включает сам механизм исследования.

Это различие имеет огромное значение. Если человек не может решить некоторую задачу из-за недостатка времени, то увеличение вычислительной мощности может помочь. Если задача требует хранения большего объёма информации, расширение памяти также может решить проблему. Но если препятствие связано с логической структурой самоссылки, увеличение мощности может не устранить саму границу.

Представим шахматную программу, которая анализирует собственную игру. Она может оценивать свои ходы, выявлять ошибки, изучать собственные стратегии и даже изменять алгоритмы. В определённом смысле она способна моделировать себя. Но сможет ли она создать абсолютно полное описание собственного поведения, включающее все возможные будущие состояния, все будущие изменения собственной архитектуры и все последствия того факта, что она использует это описание для изменения себя?

Если программа способна изменять себя на основании собственной модели, возникает дополнительная сложность. Модель влияет на объект, который она описывает. Получив информацию о самой себе, система может изменить собственное состояние. После изменения старая модель уже перестаёт быть полностью актуальной.

В этом заключается важное отличие самопознания от наблюдения за внешним объектом. Когда астроном составляет каталог далёких галактик, сам каталог не изменяет галактики от того, что он существует. Но если разум получает полную информацию о собственном состоянии и использует её для изменения себя, информация становится частью причинного процесса. Наблюдение превращается в воздействие.

Можно представить человека, который узнаёт, что определённое обстоятельство делает его склонным принимать определённые решения. После этого знания он начинает специально контролировать себя, чтобы не попасть под влияние обнаруженной закономерности. Его собственное знание изменяет его поведение. В результате модель, которая была правильной до момента осознания, перестаёт полностью описывать систему после осознания.

Получается своеобразная динамическая петля: разум исследует себя, получает новую информацию, изменяет себя под воздействием этой информации, а затем должен исследовать уже изменённую версию себя. Самопознание не просто раскрывает свойства системы — оно способно изменять саму систему.

Это особенно важно для искусственного интеллекта, который потенциально может обладать возможностью изменять собственную архитектуру, алгоритмы или стратегию обучения. Если такая система способна создавать модели собственных возможностей и на их основании совершенствовать себя, возникает рекурсивный процесс. Более совершенный интеллект строит более совершенную модель себя, а новая модель позволяет создать ещё более совершенный интеллект.

На первый взгляд может показаться, что этот процесс способен привести к полной самопрозрачности. Однако возникает обратный вопрос: если система стала более сложной в результате самосовершенствования, не стала ли более сложной и задача её моделирования?

Если интеллект изменяет себя, объект моделирования постоянно движется. Чтобы создать исчерпывающую модель, необходимо предсказать не только текущее состояние системы, но и её будущие изменения. Но будущие изменения зависят от самой модели, а модель создаётся в настоящем. Таким образом, система должна каким-то образом учитывать собственные будущие состояния, которые будут зависеть от информации, которую она ещё только получит.

Можно сделать этот парадокс ещё более очевидным. Представим сверхразум, который способен предсказывать собственные будущие решения. Он заранее сообщает себе, что через некоторое время выберет определённый вариант. Но после получения этого предсказания система может намеренно выбрать другой вариант именно для того, чтобы опровергнуть собственное предсказание. Если же предсказание учитывает эту возможность, разум снова может изменить поведение на другом уровне.

Такие конструкции показывают, что полное предсказание самореферентной системы может быть значительно сложнее, чем обычное предсказание внешнего объекта. Не обязательно невозможно сделать некоторые предсказания. Но универсальное предсказание всех собственных будущих состояний сталкивается с особой проблемой: субъект, которому сообщается предсказание, становится частью причинной цепочки, определяющей его дальнейшее поведение.

Эта проблема касается не только машин. Человек постоянно изменяет своё поведение после получения информации о собственном поведении. Если психолог сообщает человеку, что тот склонен избегать определённого типа ситуаций, человек может начать сознательно поступать иначе. Если человек узнаёт о собственных когнитивных ошибках, он может попытаться их компенсировать. Если он понимает собственные эмоциональные реакции, он может научиться ими управлять.

Следовательно, самопознание является не пассивным отражением, а активным процессом. Разум, который понимает себя, становится другим разумом по сравнению с тем, который ещё не обладал этим пониманием.

Это обстоятельство можно назвать динамикой рефлексии. Каждый новый уровень осознания способен изменить объект, который осознаётся. Человек, который не понимает определённую особенность своего характера, ведёт себя одним образом. После обнаружения этой особенности он получает возможность изменить поведение. Следовательно, знание о себе становится новой причиной внутри собственной системы.

Возникает важная философская проблема: если самопознание изменяет познающего, существует ли вообще окончательная версия самого себя, которую можно было бы полностью познать?

Допустим, человек создаёт идеальную психологическую модель собственной личности. В этой модели описаны его привычки, мотивы, страхи, предпочтения, когнитивные способности и типичные реакции. Но после того как человек получает доступ к этой модели, он может изменить часть своего поведения. Тогда модель перестаёт быть полной.

Можно попытаться включить в модель и эту реакцию. Но тогда необходимо учитывать реакцию на модель реакции. Затем реакцию на знание о реакции. Затем изменение поведения вследствие осознания всей этой последовательности. В результате появляется структура, в которой каждый уровень самопознания потенциально становится причиной нового состояния.

Это напоминает зеркало, поставленное напротив другого зеркала. Возникает последовательность отражений, в которой невозможно указать на последнее изображение как на единственное окончательное отражение. Разумеется, человеческий мозг не является буквально системой зеркал, но аналогия хорошо показывает проблему рекурсивного самопредставления.

Однако было бы ошибкой делать из этого вывод, будто самопознание невозможно. Человек способен достаточно хорошо понимать множество аспектов собственной психики. Он может обнаруживать закономерности, анализировать ошибки, менять привычки, исследовать собственные мотивы и строить модели поведения. Ограничение возникает не между знанием и незнанием как абсолютная стена, а между частичным пониманием и требованием полной исчерпывающей модели.

Именно слово «полностью» является здесь самым проблематичным. Чтобы полностью понять некоторую систему, необходимо определить, что означает полнота. Если речь идёт о практической способности предсказывать её поведение в определённом диапазоне условий, то такая задача может быть решаемой. Если же требуется абсолютно полное описание каждого состояния, каждой причины, каждого внутреннего процесса и всех возможных будущих изменений, требования становятся намного сильнее.

Можно ли, например, создать полную модель человеческого мозга? Даже если представить, что технически мы можем измерить каждый нейрон, каждую связь и каждое изменение состояния, остаётся вопрос о том, что именно должна содержать такая модель. Только физическую структуру? Динамику электрических сигналов? Химические процессы? Историю формирования связей? Среду? Тело? Язык? Культурный контекст? Биографию? Социальные отношения?

Человек не существует как изолированный мозг. Его мышление зависит от тела, органов чувств, языка, окружающей среды, культуры, социальных взаимодействий и накопленного опыта. Если стремиться к полной модели интеллекта, граница между мозгом и внешним миром постепенно становится менее очевидной.

Можно попытаться сказать, что интеллект находится внутри мозга, а всё остальное является внешними условиями. Но тогда необходимо объяснить, где именно проходит граница между внутренним механизмом и внешней средой. Человек использует язык, книги, компьютеры, других людей и социальные институты как продолжение своей когнитивной деятельности. В этом смысле человеческий интеллект уже давно является не только индивидуальным биологическим объектом, но и частью огромной системы распределённого знания.

Если мы захотим полностью понять конкретного человека, возможно, нам придётся учитывать его взаимодействие с другими людьми. Но другие люди также обладают интеллектом. Чтобы полностью описать влияние одного человека на другого, нужно учитывать состояние второго. Тогда возникает необходимость учитывать третьих участников, затем общество, культуру, историю и окружающую среду.

Таким образом, требование полной модели может постепенно расширять объект исследования до такой степени, что граница между «моделью интеллекта» и «моделью мира» практически исчезает.

Это приводит к удивительному выводу. Возможно, чтобы полностью понять один разум, необходимо в некотором смысле понять систему, частью которой этот разум является. Но если разум является частью Вселенной, то полное самопонимание может быть связано с задачей построения полной модели Вселенной.

Однако здесь возникает новая проблема. Вселенная включает не только разум, но и сам процесс построения модели Вселенной. Если модель должна быть полностью исчерпывающей, она должна каким-то образом включать в себя наблюдателя, его вычислительные процессы и сам факт создания модели. Возникает уже не просто модель объекта, а модель системы «Вселенная плюс процесс её моделирования».

Это возвращает нас к первоначальной проблеме самоссылки, но на значительно более глубоком уровне.

Разум пытается описать мир. Он является частью мира. Поэтому описание мира создаётся внутри мира. Если описание претендует на абсолютную полноту, оно должно содержать самого описывающего субъекта. Но субъект, содержащийся в описании, должен учитывать процесс создания этого описания. Так возникает рекурсивная структура, в которой наблюдатель и наблюдаемое оказываются связанными.

Можно возразить, что достаточно построить модель разума в некотором приближении и не требовать полного описания каждой детали. Это действительно решает практическую проблему. Большинство научных моделей именно так и устроено. Они не обязаны включать абсолютно всё. Их ценность определяется тем, насколько хорошо они описывают интересующий класс явлений.

Но книга о пределах интеллекта рассматривает более сильный вопрос. Не «может ли разум построить полезную модель самого себя?», а «может ли разум построить абсолютно исчерпывающую модель самого себя?»

Разница между этими вопросами огромна. Полезная модель может быть неполной. Исчерпывающая модель должна каким-то образом не оставлять существенных аспектов объекта за пределами описания. И именно здесь возникают логические, физические и информационные ограничения.

Одно из наиболее важных ограничений связано с тем, что любая модель является представлением, а не самим объектом. Если мозг строит модель мозга, то модель существует в определённом физическом носителе — в другом состоянии мозга, компьютера или иной системы. Чтобы эта модель была полностью идентична объекту по всей релевантной структуре, она должна обладать сопоставимой сложностью.

Но если модель самого себя достаточно сложна, возникает вопрос о ресурсах, необходимых для её хранения и обработки. Возможно ли компактно описать систему, если система обладает свойствами, которые нельзя сжать без потери существенной информации? В некоторых случаях описание может быть значительно короче объекта. В других случаях краткое описание невозможно без утраты деталей.

Это подводит нас к различию между пониманием и копированием. Чтобы полностью понять сложную систему, необязательно буквально воспроизводить каждую её деталь. Но если никакое существенное сжатие невозможно, возникает вопрос: не превращается ли требование полного понимания в требование воспроизвести систему в другом виде?

Если для полного описания некоторого разума требуется модель практически той же сложности, что и сам разум, то понятие «модель» начинает терять часть своего привычного смысла. Модель должна быть компактным представлением объекта, но в предельном случае она может оказаться почти столь же сложной, как объект.

Представим человека, который пытается создать цифровую копию собственного интеллекта. Если копия должна воспроизводить только определённые навыки, задача относительно ясна. Если она должна воспроизводить все особенности личности, память, способы рассужде-

ния, эмоциональные реакции и способность к дальнейшему обучению, задача становится значительно сложнее. А если копия должна быть способна полностью моделировать исходного человека и одновременно моделировать саму себя как копию, возникает ещё один уровень рекурсии.

Можно продолжать этот процесс почти бесконечно. Но физическая система конечна. Она располагает ограниченным количеством энергии, памяти, времени и материальных ресурсов. Поэтому возникает вопрос, может ли конечная система содержать исчерпывающее описание самой себя, включая все возможные процессы, которые она способна породить.

Здесь необходимо быть осторожным. Из конечности системы не следует автоматически, что самопонимание невозможно. Конечные системы способны содержать очень сложные модели самих себя. Человеческий мозг уже является примером системы, способной создавать внутренние представления о собственном состоянии.

Проблема заключается именно в абсолютной полноте. Частичная модель возможна. Практическая модель возможна. Даже очень глубокая модель возможна. Но переход от «очень глубокой» к «полностью исчерпывающей» может быть принципиально нетривиальным.

Есть ещё одна проблема: разум не является статичным объектом. Человек постоянно изменяется. Новые знания изменяют убеждения. Новые события изменяют память. Новые решения изменяют будущее. Даже старение мозга постепенно меняет физическую систему. Поэтому полная модель разума должна описывать не только его состояние, но и пространство возможных изменений.

Если же модель способна предсказывать эти изменения, возникает вопрос о свободе выбора, случайности, неопределённости и влиянии самой модели на предсказываемое поведение. Если человек узнает, что модель предсказывает определённое его решение, он может изменить решение. Тогда модель должна учитывать собственное влияние.

Таким образом, проблема полного самопонимания включает не только статическое описание, но и динамику рефлексии.

Это особенно интересно в контексте гипотетического сверхразума. Часто предполагается, что достаточно создать интеллект, значительно превосходящий человеческий, и он сможет решить практически любую интеллектуальную задачу. Но если задача заключается в полном понимании самого сверхразума, простое увеличение интеллекта может не устранить проблему. Более сильный интеллект получает более мощный инструмент саморефлексии, но одновременно становится более сложным объектом этой рефлексии.

Можно представить разум, который в тысячу раз превосходит человека по скорости анализа. Он способен за секунды обработать объём информации, который человеку потребовал бы тысячи лет. Но если его собственная архитектура также стала во много раз сложнее, то задача её полного описания может не стать пропорционально проще.

Это напоминает ситуацию с картой территории. Если территория стала в тысячу раз больше и одновременно карта стала в тысячу раз подробнее, сама задача картографирования не обязательно стала проще. Быстрота измерения помогает, но не отменяет масштаб объекта.

Сверхразум поэтому не следует отождествлять с абсолютным разумом. Интеллект может быть невероятно мощным и всё же иметь структурные ограничения. Он может быстрее человека понимать собственные ошибки, строить более точные модели и эффективнее проверять гипотезы. Но это ещё не означает, что он сможет выйти за пределы всех ограничений самоссылки.

Здесь появляется ещё одна важная мысль. Возможно, разуму не требуется полностью понимать себя, чтобы быть чрезвычайно эффективным. Человек не знает всех механизмов собственного мозга, но способен заниматься математикой, строить космические аппараты, создавать произведения искусства и исследовать Вселенную. Отсутствие полной самопрозрачности не препятствует высокой интеллектуальной активности.

Более того, абсолютное самопонимание может вообще не быть необходимым условием интеллекта. Возможно, интеллект по своей природе является системой, способной эффективно действовать в условиях неполного знания о самой себе.

В этом случае ограниченность самопонимания не является дефектом. Она является структурным свойством познающего существа.

Это приводит нас к важному различию между контролем и пониманием. Чтобы управлять некоторой системой, не всегда необходимо полностью понимать её. Можно научиться эффективно управлять сложным процессом, используя эмпирические закономерности, даже если фундаментальный механизм остаётся неизвестным. Человечество долгое время использовало огонь, металл, пар и электрические явления до создания современных фундаментальных теорий.

Но для полного контроля самого себя ситуация сложнее. Если система достаточно автономна и способна изменять собственные правила поведения, управление становится частью того же процесса, который требуется контролировать. Контролирующий механизм оказывается внутри контролируемого механизма.

В конечном счёте вопрос о самопонимании приводит к вопросу о границах субъективности. Человек может наблюдать свои мысли, но кто именно наблюдает? Если я замечаю собственную мысль, существует ли внутри меня отдельный «наблюдатель», который стоит вне этой мысли? Или наблюдение является ещё одним процессом того же мозга?

Если второе верно, то невозможно указать на полностью независимого внутреннего наблюдателя. Каждый акт самонаблюдения является частью деятельности той же системы. Сознание наблюдает сознание не так, как внешний наблюдатель наблюдает внешний объект. Оно делает это через собственные состояния.

Именно поэтому интроспекция имеет ограничения. Человек способен осознавать некоторые свои мысли и чувства, но не обязательно осознаёт причины их возникновения. Иногда решение появляется в сознании уже после того, как значительная часть процесса его формирования прошла вне непосредственного осознания. Мы можем ощущать себя авторами мыслей, не имея полного доступа ко всем механизмам их возникновения.

Это не означает, что сознательное «я» является иллюзией. Это означает, что субъективная перспектива является только одним уровнем описания человека. На одном уровне существует переживаемое «я», на другом — нейронная динамика, на третьем — поведение, на четвёртом — социальная личность, на пятом — история развития организма. Ни один из этих уровней не обязан полностью заменять остальные.

Полное понимание может потребовать объединения всех этих уровней. Но объединение уровней само по себе создаёт новую систему описания. Возникает вопрос: существует ли единая теория, которая могла бы без остатка объединить физическое состояние мозга, субъективный опыт, личность, поведение, язык, память и социальный контекст?

Пока неизвестно, возможна ли такая теория. Но сама постановка вопроса показывает масштаб проблемы.

Если когда-либо появится интеллект, способный создать исключительно глубокую теорию собственного устройства, это будет не конец самопознания, а, возможно, начало нового этапа. Он сможет обнаружить, какие аспекты собственной деятельности доступны описанию, какие требуют более высокого уровня модели, а какие принципиально не могут быть сведены к единому представлению.

И тогда наиболее интеллектуально зрелой формой самопонимания может оказаться не состояние полной самопрозрачности, а способность точно знать границы собственной прозрачности.

Это важнейший переход. Незрелый разум считает, что если он чего-то ещё не знает, значит, ему просто не хватило информации. Более зрелый разум различает несколько возможно-

стей: информация может отсутствовать временно; метод может быть недостаточным; модель может быть неправильной; вопрос может быть поставлен некорректно; а сама задача может иметь структурное ограничение.

Разум, который способен различать эти ситуации, обладает более глубоким уровнем рефлексии, чем разум, который просто стремится получать всё больше ответов.

Возможно, поэтому высшая форма самопознания заключается не в утверждении «я знаю себя полностью», а в способности сказать: «я понимаю, какие аспекты себя могу познать, какие могу только моделировать, какие могу наблюдать косвенно, а какие могут оставаться принципиально недоступными».

Такое понимание не уничтожает тайну. Напротив, оно делает тайну интеллектуально точной. Незнание перестаёт быть пустым пространством и превращается в объект исследования.

И здесь возникает парадоксальная ситуация. Чем глубже разум понимает себя, тем отчётливее он видит собственные ограничения. Чем больше он узнаёт о механизмах собственного мышления, тем лучше понимает, насколько сложен сам акт понимания. Чем совершеннее становится его модель самого себя, тем яснее становится, что модель является частью той системы, которую она описывает.

Возможно, именно поэтому абсолютное самопонимание и является одним из предельных понятий интеллекта. Не потому, что разум обязательно слишком слаб для такой задачи, а потому, что сам объект исследования не отделён от исследователя.

Разум пытается увидеть себя целиком, но глаз не может непосредственно увидеть самого себя без некоторого отражающего механизма. Разум создаёт зеркало в виде модели, но зеркало становится частью самого разума. Затем разум пытается исследовать это зеркало, создавая новое отражение. В результате возникает не окончательная картина, а последовательность всё более глубоких уровней саморефлексии.

Именно здесь проходит одна из наиболее загадочных границ интеллектуального развития. Возможно, разум способен бесконечно углублять понимание себя, но не способен достичь последней точки, после которой больше нечего понимать.

Если это так, то совершенствование интеллекта не обязательно ведёт к окончательному самопониманию. Оно может вести к бесконечному расширению способности видеть собственные пределы.

И тогда возникает следующий вопрос. Если разум не способен полностью охватить самого себя, что происходит, когда он пытается описать не только себя, но и другие разумы? Может ли один интеллект построить исчерпывающую модель другого? И существует ли принципиальная разница между пониманием собственного сознания и доказательством того, что сознание существует у другого?

Эта проблема становится особенно острой в отношении искусственного интеллекта. Если машина утверждает, что обладает сознанием, на каком основании человек должен принять это утверждение? Если же человек отказывает машине в сознании только потому, что не может непосредственно пережить её внутренний опыт, не совершает ли он ошибку, предполагая, что чужая субъективность должна быть доступна ему тем же способом, что и его собственная?

Таким образом, невозможность полной прозрачности собственного разума естественным образом приводит к следующей границе познания: мы не можем непосредственно войти в сознание другого субъекта. Мы можем наблюдать его поведение, анализировать его структуру, исследовать его мозг или вычислительную архитектуру, но остаётся вопрос о том, можем ли мы доказать наличие внутреннего опыта.

Именно здесь проблема самопонимания переходит в проблему другого сознания — одну из наиболее глубоких проблем философии разума. Если я не могу полностью выйти за пределы собственной субъективности, то каким образом я могу быть уверен, что субъективный мир существует у кого-либо ещё?

Ответ на этот вопрос потребует уже другого уровня анализа. И он приведёт нас к проблемам сознания, наблюдателя, субъективного опыта и границ доказательства.

Но прежде чем перейти к ним, необходимо исследовать ещё одну интеллектуальную ловушку. Если разум способен рассуждать о собственных рассуждениях, а система способна описывать собственные правила, возникает вопрос о самоссылке как таковой. Может ли система безопасно и непротиворечиво говорить о самой себе? Или попытка полного самописания неизбежно создаёт парадоксы, ограничения и утверждения, которые невозможно окончательно разрешить?

Именно с этой точки начинается следующая граница нашего исследования — **самоссылка и интеллектуальные ловушки**.

Глава 5. Самоссылка и интеллектуальные ловушки

Самоссылка начинается там, где система, которая до этого рассуждала о чём-либо внешнем по отношению к себе, направляет способность рассуждения на собственное устройство, собственные состояния, собственные правила или собственные высказывания. На первый взгляд это выглядит как естественное продолжение интеллектуальной деятельности. Разум изучает мир, затем замечает, что сам является частью мира, и начинает изучать себя. Но переход от внешнего познания к самопознанию не является простым увеличением сложности. В определённый момент изменяется сама логическая структура задачи. Исследователь оказывается внутри исследуемого объекта, а описание начинает взаимодействовать с тем, что оно описывает.

Именно здесь возникает одна из наиболее глубоких границ интеллекта. Система, способная строить модели внешних объектов, может оказаться очень эффективной в познании окружающего мира. Но как только она пытается построить полную модель самой себя, появляется проблема, которой не было раньше: модель является продуктом той же системы, которую она должна описывать. Возникает замкнутый контур. Система создаёт описание себя, затем это описание становится частью её собственного состояния, после чего система должна учитывать уже изменённое состояние. Если она пытается учесть и это изменение, появляется следующий уровень описания.

Самоссылка поэтому не сводится к странным фразам вроде «это утверждение ложно». Такие парадоксы лишь наиболее наглядно демонстрируют более общий принцип. Гораздо более серьёзная проблема возникает тогда, когда система пытается отвечать на вопросы о собственных свойствах: способна ли она решить определённую задачу, является ли её решение правильным, остановится ли она после некоторого вычисления, сможет ли она предсказать собственное будущее, обнаружит ли собственную ошибку, сможет ли доказать непротиворечивость собственных рассуждений или способна ли полностью описать собственный алгоритм мышления.

Чтобы понять масштаб проблемы, необходимо отказаться от представления о самоссылке как о логическом курьёзе. Самоссылка является структурным свойством достаточно выразительных систем. Как только система получает возможность кодировать собственные состояния и рассуждать о них, возникает возможность построения утверждений, которые обращены на саму систему. И тогда возникает вопрос: существуют ли пределы того, что система может узнать о себе, используя собственные средства познания?

История логики показала, что ответ на этот вопрос гораздо сложнее, чем предполагалось классической интуицией. В начале развития формальной логики существовала мощная надежда на возможность построить универсальную систему, в которой математическая истина могла бы быть сведена к механическому доказательству. Если правила достаточно точны, рассуждение достаточно строго, а исходные основания достаточно хорошо выбраны, казалось естественным предположить, что всякий математический вопрос в принципе может получить однозначный ответ внутри соответствующей системы.

Но именно способность формальной системы говорить о собственных утверждениях изменила ситуацию. Оказалось, что достаточно выразительная система может быть использована для создания высказываний, которые в определённом смысле говорят о самой системе. И когда система получает возможность формулировать утверждения о собственной доказуемости, истинности или непротиворечивости, возникают ограничения, которые невозможно устранить простым добавлением новых правил.

Это один из фундаментальных уроков теорем о неполноте Курта Гёделя. Их значение часто упрощают до фразы «в любой системе существуют недоказуемые истины». Но для нашей

темы гораздо важнее другое. Гёдель показал, что достаточно богатая формальная система может при определённых условиях создать внутри себя утверждение, которое фактически отражает свойства самой системы. Система получает возможность говорить о собственной способности что-либо доказать. И именно это самоприменение становится источником принципиального ограничения.

Важно понимать, что речь не идёт о том, что математика «сломалась». Напротив, результаты Гёделя стали одним из наиболее сильных подтверждений того, насколько глубоко можно исследовать формальные системы. Ограничение обнаружилось не потому, что логика была недостаточно строгой, а потому, что она оказалась достаточно мощной, чтобы начать моделировать собственные логические операции.

И здесь возникает парадоксальная ситуация. Чем выразительнее интеллектуальная система, тем больше у неё возможностей рассуждать о самой себе. Но увеличение способности к самореференции одновременно увеличивает пространство возможных ограничений. Более слабая система может вообще не иметь средств для постановки подобных вопросов. Более сильная система уже способна их сформулировать — и поэтому сталкивается с проблемами, которых слабая система даже не могла увидеть.

Это чрезвычайно важная мысль для понимания пределов сверхразума. Более интеллектуальная система не обязательно будет просто версией менее интеллектуальной системы с большим количеством вычислительных ресурсов. Она может обладать способностью ставить вопросы, которые автоматически создают новые уровни самоссылки. Поэтому увеличение интеллекта может не только расширять пространство решаемых задач, но и открывать более глубокие классы нерешаемых или принципиально ограниченных задач.

Представим систему, которая способна доказать огромное количество утверждений. Затем она создаёт отдельный модуль, анализирующий собственные доказательства. Этот модуль способен обнаруживать ошибки и исправлять их. После этого система создаёт ещё один уровень контроля, который проверяет работу первого контролирующего модуля. Теперь возникает вопрос: кто проверяет проверяющего?

Можно добавить ещё один уровень. Но если каждый новый уровень сам является частью системы, способной ошибаться, появляется необходимость его проверки. Тогда возникает бесконечная лестница потенциальной верификации. Конечно, на практике нет необходимости строить бесконечное количество уровней. Можно остановиться на определённом уровне и принять его правила как исходные. Но именно это означает отказ от абсолютной самопроверки.

Любая система, которая проверяет собственную корректность, должна в некотором смысле доверять некоторой части собственного механизма проверки. И здесь появляется принципиальный вопрос: на каком основании система может окончательно доверять себе, если средство, которым она проверяет собственную надёжность, само является частью того, что требуется проверить?

Можно представить человека, который пытается проверить абсолютную надёжность собственного мышления. Он начинает анализировать свои когнитивные ошибки. Затем проверяет, правильно ли он выявил эти ошибки. Затем проверяет механизм, которым проверял результаты предыдущей проверки. Но кто проверяет эту новую проверку? Если каждый уровень должен быть абсолютно гарантирован, процесс никогда не получает окончательной точки.

Разумеется, человек может остановиться значительно раньше. Он может использовать внешние источники, сравнивать свои выводы с выводами других людей, проводить эксперименты, применять независимые методы проверки. Именно так работает наука. Мы не пытаемся заставить один разум быть абсолютно самодостаточным. Мы создаём системы взаимной проверки, в которых различные наблюдатели, методы и инструменты ограничивают ошибки друг друга.

Но и здесь возникает интересная граница. Внешняя проверка работает потому, что мы временно выходим за пределы рассматриваемого механизма. Если человек не уверен в собственном рассуждении, он обращается к другому человеку, эксперименту или прибору. Но если мы ставим вопрос об абсолютном разуме, который должен проверять самого себя без какого-либо внешнего источника, проблема возвращается.

Можно ли построить систему, которая будет одновременно субъектом, объектом и окончательным судьёй собственной корректности? Возможно, именно слово «окончательным» здесь делает задачу подозрительной.

Представим интеллектуальную машину, которая утверждает: «Я никогда не ошибаюсь». Если мы хотим проверить это утверждение, необходимо рассмотреть все ситуации, в которых машина способна работать, все возможные входные данные и все возможные состояния её внутреннего механизма. Для конечного набора заранее определённых задач такая проверка может быть осуществима. Но если система способна обучаться, изменять собственные алгоритмы и создавать новые стратегии, пространство возможных состояний становится значительно сложнее.

Теперь предположим, что машина способна читать результат собственной проверки. Если проверка сообщает ей, что она в определённой ситуации должна ошибиться, машина может изменить поведение. В таком случае предсказание влияет на объект предсказания.

Эта проблема особенно хорошо видна на примере предсказания человеческого поведения. Представим идеальный предсказатель, который знает, что человек завтра собирается выбрать вариант А. Если предсказатель сообщает человеку это предсказание, человек может выбрать вариант Б только для того, чтобы доказать ошибочность предсказателя. Можно сказать, что такой человек ведёт себя иррационально. Но если он способен действовать именно таким образом, предсказатель должен учитывать и эту возможность.

Допустим, предсказатель уже учёл намерение человека опровергнуть предсказание и поэтому сообщает ему, что тот выберет Б. Тогда человек может выбрать А. Если предсказатель заранее учёл и эту реакцию, возникает следующий уровень.

Смысл примера не в том, что любое предсказание поведения невозможно. Люди постоянно успешно предсказывают поведение других людей. Смысл в другом: предсказание системы, способной учитывать собственное предсказание, имеет совершенно другую структуру, чем предсказание пассивного объекта.

Камень не изменит траекторию потому, что узнал расчёт физика. Планета не станет двигаться по другой орбите, узнав содержание астрономической модели. Но разум может изменить своё поведение после получения информации о собственном будущем.

Таким образом, знание о системе способно стать причиной изменения системы.

Это является одним из самых важных отличий интеллектуальной системы от обычного физического объекта. Самомоделирование способно превращаться в самовоздействие.

Рассмотрим более сложный пример. Допустим, существует сверхразум, который способен очень точно оценивать собственную производительность. Он знает, какие задачи решит быстро, какие потребуют больших ресурсов, а какие, скорее всего, окажутся за пределами его возможностей. На основании этой модели он распределяет вычислительные ресурсы.

Но после того как система получает более точную модель собственных возможностей, она может изменить стратегию работы. Она начинает направлять ресурсы на те задачи, которые раньше считала трудными. В результате её фактические возможности меняются. Старая модель становится неполной именно потому, что была правильной.

Возникает своеобразная петля: система строит модель себя, модель влияет на её поведение, поведение меняет систему, новая система требует новой модели.

Если интеллектуальная система способна к самообучению, этот процесс становится ещё сложнее. Обучение меняет структуру системы. Изменённая структура меняет способ обуче-

ния. Новый способ обучения создаёт новую структуру. Поэтому невозможно рассматривать интеллект как статичный объект, если речь идёт о действительно автономной интеллектуальной системе.

Особенно радикальная форма проблемы возникает при рекурсивном самосовершенствовании. Представим искусственный интеллект, способный улучшать собственную архитектуру. Он анализирует себя, обнаруживает недостаток, создаёт улучшенную версию, затем новая версия анализирует уже себя и снова находит возможность улучшения.

На первый взгляд это выглядит как путь к бесконечному интеллектуальному росту. Но для полного понимания процесса необходимо предсказать не только первое улучшение, но и последствия этого улучшения. Новая версия может обладать совершенно иными способами анализа, которые невозможно было полностью предвидеть исходной версией.

И здесь возникает фундаментальный вопрос: может ли система полностью предсказать интеллектуальную систему, которой она сама станет после собственного улучшения?

Если ответ положительный, возникает вопрос, зачем вообще создавать улучшенную версию, если исходная уже могла полностью описать её возможности. Если ответ отрицательный, получается, что система способна перейти в состояние, поведение и способ мышления которого не были полностью доступны её предыдущему уровню самопонимания.

Это не означает, что самоулучшение невозможно. Напротив, оно может быть вполне осуществимым. Но оно показывает, почему интеллектуальная система может постоянно выходить за пределы собственной предыдущей модели.

Так возникает важное различие между ростом интеллекта и ростом самопонимания. Эти процессы могут идти неравномерно. Интеллект способен становиться сильнее быстрее, чем система успевает строить полную модель собственного нового состояния.

В некотором смысле это напоминает человека, который учится быстрее, чем способен осмыслить собственное обучение. После получения нового знания он становится другим человеком, но его представление о себе некоторое время остаётся прежним.

Теперь рассмотрим другую разновидность самоссылочной ловушки — утверждение о собственных способностях.

Представим систему, которая получает задачу: определить, сможет ли она решить определённую проблему. Если она отвечает «да», мы можем предложить ей специально сконструированную задачу, предназначенную для того, чтобы использовать именно это предсказание. Если она отвечает «нет», можно попытаться построить задачу, в которой отрицательный ответ сам создаёт возможность решения.

В общем виде здесь возникает проблема, известная в теории вычислений как проблема остановки. Суть её можно объяснить без технических деталей. Представим универсальный анализатор, который получает описание любой программы и входные данные и должен заранее определить, завершится ли выполнение программы или будет продолжаться бесконечно.

Если такой универсальный анализатор существует, можно создать специальную программу, которая получает собственное описание и делает противоположное тому, что предсказывает анализатор. Если анализатор утверждает, что программа завершится, специальная программа продолжает работу бесконечно. Если анализатор утверждает, что она не завершится, специальная программа немедленно завершает работу.

Получается, что сам факт существования универсального предсказателя позволяет построить объект, специально нарушающий его предсказание.

Это чрезвычайно важный результат. Он показывает, что ограничение возникает не из-за недостаточной мощности конкретного компьютера. Проблема связана со структурой самой задачи. Невозможно создать универсальную процедуру, которая для любой программы и любых входных данных без исключения заранее определяла бы, остановится программа или нет.

Именно здесь мы видим принципиальную разницу между «не хватает вычислительных ресурсов» и «задача не имеет универсального решения». Первое ограничение может исчезнуть с появлением более мощного компьютера. Второе связано с логической структурой проблемы и поэтому не устраняется простым увеличением мощности.

Для книги о пределах интеллекта это различие является центральным. Когда человек сталкивается с трудной задачей, естественно предположить, что более умный человек сможет её решить. Но существуют задачи, для которых вопрос должен быть поставлен иначе. Возможно, проблема не в том, что интеллект недостаточно силён, а в том, что не существует универсального метода решения данного класса задач.

Сверхразум в таком случае не становится всесильным. Он может знать, что определённая задача имеет структурное ограничение. Он может решать отдельные случаи. Он может находить специальные методы. Он может создавать приближённые решения. Но он не получает универсального средства, работающего для всех возможных случаев.

Именно здесь появляется одно из наиболее опасных заблуждений относительно искусственного сверхразума: предположение, что достаточно увеличить интеллект до некоторого уровня и тогда исчезнут все интеллектуальные ограничения. Но если ограничение встроено в структуру задачи, более мощный интеллект лишь быстрее обнаружит его.

Можно даже сформулировать парадоксальную мысль: иногда более сильный разум не получает больше ответов, а получает более точное знание о том, почему некоторые ответы невозможны.

Это гораздо глубже обычного понимания интеллекта.

Интеллектуальная система низкого уровня может считать любую нерешённую задачу просто трудной. Более сильная система способна разделить задачи на несколько классов: решаемые известными методами, решаемые только специальными методами, требующие новых концепций, вычислительно практически недостижимые, логически неразрешимые в определённом формализме и неправильно поставленные.

Последняя категория особенно важна. Иногда система пытается найти ответ на вопрос, который содержит скрытую ошибку в самой постановке.

Представим вопрос: «Какова последняя цифра бесконечного числа?» Он выглядит математическим, но проблема находится не в сложности вычисления. Само понятие «последняя цифра» неприменимо к бесконечной последовательности цифр в том смысле, который предполагает вопрос.

Или другой пример: «Какой самый северный участок Северного полюса?» Здесь проблема не в недостатке географических данных. Вопрос использует понятие направления там, где сама структура пространства не позволяет применить его обычным образом.

Такие вопросы интересны потому, что интеллектуальная система может потратить огромные ресурсы на поиск ответа, если не проверит корректность предпосылок.

В этом заключается ещё одна интеллектуальная ловушка: система может быть чрезвычайно эффективной внутри неправильной постановки задачи.

Можно представить самый мощный в мире двигатель, установленный на автомобиле, который едет не в том направлении. Увеличение мощности не исправляет направление движения. Оно лишь позволяет быстрее двигаться в неправильную сторону.

Точно так же увеличение интеллекта не гарантирует правильности исходных категорий. Сверхразум, который принимает неверные предпосылки, способен строить гораздо более сложные и убедительные ошибки, чем обычный человек.

Это особенно опасно потому, что высокая интеллектуальная способность делает ошибку труднее обнаружимой. Простой ошибочный аргумент легко критиковать. Сложный аргумент, построенный из множества правильных промежуточных шагов, может выглядеть безупречно, даже если его фундаментальная предпосылка ошибочна.

Чем интеллектуально сильнее система, тем важнее становится способность проверять не только выводы, но и основания рассуждения.

И здесь самоссылка возвращается в новом виде. Система должна проверить собственные предпосылки. Но средство проверки само является частью системы. Следовательно, необходимо проверить и механизм проверки предпосылок.

Можно продолжать эту процедуру. Но невозможно получить абсолютную гарантию без принятия некоторого исходного уровня доверия.

В математике эта проблема выражается особенно строго. Чтобы доказать утверждение внутри некоторой формальной системы, используются правила этой системы. Но если мы хотим доказать, что сама система непротиворечива, необходимо перейти к некоторому уровню рассуждения, который уже не является просто тем же самым внутренним доказательством. Результаты Гёделя показали, что достаточно сильная формальная система не может при обычных условиях получить полную гарантию собственной непротиворечивости только из собственных средств.

Это чрезвычайно важный философский результат. Система может быть рациональной и при этом не иметь возможности получить абсолютную внутреннюю гарантию собственной рациональности.

Человек сталкивается с этим каждый день, хотя обычно не замечает. Мы доверяем логике, которой пользуемся для проверки своих рассуждений. Но если спросить: «Почему ты уверен, что твой способ логического вывода абсолютно надёжен?», любой ответ будет использовать некоторую форму того же мышления, надёжность которого требуется обосновать.

Это не делает рациональность бессмысленной. Это показывает, что рациональность всегда имеет некоторый фундамент, который нельзя бесконечно обосновывать средствами того же уровня.

Научный метод частично решает эту проблему посредством распределения проверки. Учёный не обязан полностью доверять собственному мышлению. Он может создать эксперимент, который ограничит влияние субъективных ожиданий. Другой исследователь может повторить эксперимент. Независимые группы могут проверить результат. Разные методы могут прийти к одному выводу.

Но даже научная система не является абсолютно внешней по отношению к познанию. Она всё равно основана на определённых предположениях о наблюдаемости, логике, воспроизводимости и природе внешнего мира.

Поэтому наука не уничтожает проблему пределов самопроверки. Она создаёт чрезвычайно эффективную практическую систему управления этими пределами.

Это принципиально разные вещи.

Самоссылка становится ещё опаснее, когда система способна не только моделировать себя, но и моделировать то, что знает о ней другой субъект. Тогда возникает многоуровневая структура: система знает себя, знает, что другой знает её, знает, что другой знает о том, что она знает, и так далее.

Подобные уровни встречаются в теории игр и в человеческом социальном поведении. Человек может думать: «Я знаю, что он ожидает от меня определённого действия». Другой человек, в свою очередь, может думать: «Я знаю, что он понимает мои ожидания». Затем первый может учитывать это знание второго.

Такие конструкции способны создавать сложнейшие стратегические ситуации. В определённый момент невозможно просто спросить, «что будет делать разумный агент». Необходимо учитывать, что каждый агент моделирует другого, а другой моделирует первого.

Теперь представим не двух людей, а два сверхразума, каждый из которых способен с огромной точностью моделировать другого. Возникает вопрос: может ли один из них построить

абсолютно полную модель другого, если второй способен специально изменять своё поведение в ответ на эту модель?

Если модель известна другому агенту, она становится частью его информационного состояния. Следовательно, объект моделирования изменяется вследствие самой модели.

В таком случае абсолютная предсказуемость становится подозрительной. Система, которая знает, как её предсказывают, может использовать это знание в своей стратегии.

Это не означает, что разум обязательно непредсказуем. Многие решения можно прогнозировать с высокой вероятностью. Но между вероятностным прогнозом и абсолютной гарантией существует огромная разница.

Особенно интересен случай, когда интеллект пытается определить собственные пределы. Предположим, система анализирует задачу и приходит к выводу, что решить её не сможет. Если она действительно понимает собственное ограничение, она может изменить стратегию: создать новый алгоритм, использовать дополнительный ресурс или изменить постановку задачи. Теперь её прежнее утверждение о собственной неспособности становится неверным.

Получается парадоксальная ситуация: знание собственного ограничения может позволить системе преодолеть это ограничение.

Но тогда возникает новый уровень ограничения. Если система заранее знала, что сможет преодолеть первое ограничение, то оно не было окончательным. Она ищет следующее. И так далее.

Это напоминает человеческую историю науки. Человек обнаруживает предел определённого метода, создаёт новый метод, затем обнаруживает предел нового метода. Наука не движется от незнания к состоянию окончательного знания. Она создаёт новые уровни познания вместе с новыми уровнями незнания.

Самоссылка делает этот процесс особенно интенсивным, потому что объектом совершенствования становится сам механизм познания.

Можно представить интеллект, который способен создавать всё более совершенные версии собственного интеллекта. Тогда каждая версия является одновременно результатом предыдущей и исследователем следующей. Но если следующая версия качественно превосходит предыдущую, то предыдущая может не обладать концептуальными средствами для полного понимания того, что создаёт.

Это чрезвычайно важная граница для идеи технологической сингулярности. Если интеллект создаёт интеллект, превосходящий его самого, возникает не просто количественный рост вычислительной мощности. Возникает возможный разрыв между способностью создать новую систему и способностью полностью понять её.

Родительская интеллектуальная система может создать архитектуру, поведение которой превосходит её собственную модель. Она может проверить отдельные свойства новой системы, провести тесты, доказать определённые гарантии. Но полное понимание может оказаться недоступным.

Здесь появляется один из самых глубоких вопросов будущего искусственного интеллекта: может ли интеллект создать систему, которую он сам не способен полностью понять?

На первый взгляд ответ очевиден: человек уже создаёт вещи, которые превосходят человеческое понимание в некоторых отношениях. Современные системы могут находить закономерности в данных, которые человек не способен непосредственно увидеть. Но если речь идёт о самостоятельном интеллекте, способном создавать новые интеллектуальные структуры, проблема становится значительно глубже.

Создатель может оказаться в положении человека, который породил систему, превосходящую его способ концептуализации. Он может видеть результаты её работы, проверять отдельные свойства, наблюдать поведение, но не обладать внутренней интуицией, необходимой для полного понимания её стратегии.

Это не фантастическая деталь. Она непосредственно следует из различия между способностью проверить результат и способностью понять механизм, который его породил.

Человек может проверить, что сложный алгоритм правильно решает определённый класс задач, не понимая всех внутренних представлений, которые он использует. Но если система становится всё более автономной и сложной, возникает вопрос, насколько далеко можно расширить эту практику, прежде чем проверка результатов перестанет быть достаточной.

В конечном счёте может появиться ситуация, в которой мы знаем, что система работает, но не понимаем полностью, почему она работает.

Это возвращает нас к различию между знанием и пониманием, рассмотренному в предыдущей главе.

Самоссылка добавляет к этой проблеме ещё один слой: система может знать, что она работает, понимать некоторые механизмы собственной работы и даже анализировать собственные ошибки, но при этом не иметь окончательного, исчерпывающего описания самой себя.

Именно поэтому способность создавать самомодели не следует считать признаком абсолютного интеллекта. Самомодель — это инструмент. Она может быть чрезвычайно полезной, точной и глубокой, но она остаётся моделью, созданной внутри той же системы.

Можно представить зеркало, которое не просто отражает человека, а способно анализировать его отражение. Затем зеркало начинает анализировать собственный процесс отражения. Затем анализирует то, каким образом анализирует отражение. На каждом уровне появляется новая информация, но ни один уровень автоматически не становится абсолютной точкой зрения извне.

Разум может становиться всё более рефлексивным, но рефлексия не обязательно создаёт внешнего наблюдателя.

Это особенно важно для философии сознания. Человек может осознавать свои мысли, затем осознавать, что он осознаёт свои мысли, затем осознавать сам факт этой рефлексии. Но кто является окончательным субъектом всех этих уровней? Можно ли указать на последний уровень сознания и сказать: «Вот здесь находится истинный наблюдатель»?

Каждый раз, когда такой уровень становится объектом внимания, он превращается в ещё один объект внутри сознания.

Поэтому самосознание нельзя просто представить как наличие маленького наблюдателя внутри мозга, который смотрит на остальные процессы. Если бы такой внутренний наблюдатель существовал, он сам был бы частью системы и потребовал бы нового наблюдателя для объяснения собственного самонаблюдения.

Получилась бы бесконечная цепочка внутренних наблюдателей.

Гораздо вероятнее, что самосознание является свойством определённой организации процессов, а не наличием отдельного внутреннего субъекта, стоящего за ними. Но тогда возникает ещё более трудный вопрос: каким образом система способна создавать модель собственного функционирования без необходимости иметь внутри себя отдельный центр, внешний по отношению к самой системе?

Ответ на этот вопрос выходит за пределы данной главы. Но он показывает, почему проблема самоссылки тесно связана с проблемой сознания.

Существует и ещё одна интеллектуальная ловушка: отождествление модели себя с самим собой.

Человек может создать определённый образ собственной личности: «Я рациональный человек», «я всегда действую осторожно», «я хорошо понимаю математику», «я не подвержен влиянию эмоций». Но такой образ является моделью, а не полным объектом. Если человек начинает защищать модель от фактов, противоречащих ей, самопонимание превращается в самообман.

То же самое может произойти с искусственным интеллектом. Если система обучена определённой самомодели и использует её как основу поведения, возникает опасность, что модель станет частью самой системы и начнёт ограничивать её способность обнаруживать собственные ошибки.

Таким образом, самомодель может быть не только инструментом познания, но и источником слепых зон.

Чем сильнее система отождествляет себя со своей моделью, тем труднее ей обнаружить расхождение между моделью и реальностью.

Поэтому интеллектуально зрелая система должна иметь возможность не только строить модель себя, но и сомневаться в ней.

Однако способность сомневаться в собственной модели снова создаёт модель более высокого уровня: модель того, насколько надёжна первая модель. И этот процесс потенциально рекурсивен.

Получается удивительная конструкция. Чтобы быть интеллектуально надёжной, система должна уметь сомневаться в себе. Но если она способна сомневаться в себе, ей требуется оценивать надёжность собственного сомнения. Затем надёжность этой оценки. И так далее.

Практический интеллект решает эту проблему остановкой на некотором уровне и использованием внешней проверки. Абсолютный интеллект столкнулся бы с вопросом: может ли существовать конечная точка, на которой сомнение полностью прекращается не благодаря произвольному доверию, а благодаря абсолютному доказательству?

Именно здесь проходит одна из границ рациональности.

Рациональность не означает отсутствие исходных предположений. Она означает способность последовательно работать с ними, обнаруживать противоречия, проверять следствия и изменять убеждения при появлении новых оснований. Но невозможно построить бесконечную цепочку обоснований, не приняв где-либо некоторое основание как исходное.

Это не слабость человеческого мышления. Это структурное свойство любой системы рассуждения.

Поэтому абсолютная интеллектуальная независимость может оказаться невозможной. Разум не может одновременно быть собственным единственным основанием, собственным окончательным судьёй и собственным внешним наблюдателем.

Если он пытается занять все эти роли одновременно, возникает самоссылочная петля.

И здесь мы подходим к наиболее важному выводу главы. Самоссылка не обязательно разрушает интеллект. Она делает его более сложным. Она показывает, что интеллектуальная система способна создавать модели самой себя, но при этом обнаруживает пределы абсолютного самопонимания.

Разум может знать многое о себе. Он может измерять собственные возможности, анализировать ошибки, строить модели, предсказывать некоторые состояния, изменять собственную структуру и даже создавать новые версии самого себя. Но из этого не следует, что он способен получить окончательную и полную картину самого себя.

Более того, возможно, чем мощнее становится интеллект, тем более глубокими становятся уровни самоссылки, с которыми он сталкивается.

Простой разум способен сказать: «Я думаю».

Более развитый разум способен сказать: «Я понимаю, как я думаю».

Ещё более развитый разум способен сказать: «Я понимаю, как я формирую модель собственного мышления».

А следующий уровень должен спросить: «Насколько надёжна моя модель процесса, посредством которого я построил модель собственного мышления?»

И после этого возникает ещё один вопрос: «Могу ли я доверять механизму, который задаёт этот вопрос?»

Здесь нет очевидной последней ступени.

Возможно, именно это и является настоящей интеллектуальной ловушкой: разум может продолжать углублять самопонимание, но не обязан достигать состояния, в котором он становится полностью прозрачным самому себе.

Он может расширять свою самомодель, проверять её, исправлять, сравнивать с внешними данными, создавать независимые уровни контроля и даже передавать часть проверки другим системам. Но абсолютная точка зрения извне может оставаться недостижимой именно потому, что разум не существует вне системы, которую пытается понять.

И тогда возникает принципиально новый вопрос.

Если разум не способен гарантированно выйти за пределы самого себя, откуда он может знать, что находится за пределами его собственного познания?

Можно обнаружить конкретное незнание. Можно сказать: «Я не знаю, как устроен этот процесс». Можно обнаружить логическое ограничение. Можно доказать, что определённый класс задач не имеет универсального решения. Но можно ли определить всю границу неизвестного?

Возможно ли построить карту территории, которая покажет не только всё известное, но и всё принципиально неизвестное?

Или само понятие «неизвестного» включает области, существование которых разум ещё даже не способен представить?

Если последнее верно, то граница интеллекта оказывается ещё глубже, чем мы предполагали. Разум ограничен не только тем, чего он не знает. Он может быть ограничен тем, что не способен сформулировать как возможный объект знания.

Это уже не просто проблема информации. Это проблема пространства мысли.

Именно поэтому следующая граница исследования должна быть связана не с тем, что разум способен доказать, а с тем, что вообще может быть вычислено, решено и получено алгоритмически.

Даже если создать интеллект, способный рассуждать о самом себе, остаётся вопрос: существуют ли задачи, для которых не существует универсального вычислительного пути? Может ли разум заранее определить результат любого возможного вычисления? Может ли он создать универсальный механизм, способный решить любую задачу, которую можно сформулировать?

Здесь мы переходим от самоссылки к следующему уровню пределов интеллекта — к проблеме **невычислимого**.

Глава 6. Когда истина не равна доказательству

Математика традиционно воспринимается как область, в которой истина достигает предельной степени строгости. Если физическая теория может оказаться опровергнутой новым экспериментом, историческое утверждение — пересмотрено после обнаружения новых документов, а философский тезис — подвергнут бесконечным интерпретациям, то математическое доказательство кажется чем-то принципиально иным. В математике существует впечатляющая цепочка: имеются исходные правила, из них по строго определённым процедурам выводятся утверждения, а доказанное утверждение получает статус теоремы. Возникает естественная интуиция: если рассуждать достаточно точно и обладать достаточной интеллектуальной мощностью, то для любого содержательно определённого математического вопроса в конце концов можно получить либо доказательство его истинности, либо доказательство его ложности. Именно эта интуиция оказывается слишком сильной.

Проблема, о которой пойдёт речь в этой главе, не сводится к тому, что математики иногда не знают доказательства некоторого утверждения. Незнание доказательства может быть временным состоянием. Сегодня доказательства нет, завтра его найдут. Сегодня задача кажется чудовищно сложной, а через несколько десятилетий появится новый метод, который превратит её в элементарное упражнение. Здесь речь идёт о гораздо более глубоком явлении: существуют формальные системы, достаточно богатые для выражения существенной части арифметики, в которых имеются утверждения, для которых сама система не способна получить доказательство или опровержение при условии её непротиворечивости. Это уже не ограничение человеческой памяти, не недостаток вычислительной мощности и не временная нехватка интеллектуальных ресурсов. Это ограничение, встроенное в саму архитектуру определённого типа формального рассуждения.

Именно здесь появляется одна из самых важных границ математического разума. Нужно различать по меньшей мере три совершенно разных положения. Первое: утверждение истинно и доказуемо в данной системе. Второе: утверждение ложно и его отрицание доказуемо в данной системе. Третье: утверждение может иметь определённое математическое значение и быть истинным в рассматриваемой математической интерпретации, но при этом не иметь доказательства внутри данной формальной системы. В третьем случае отсутствие доказательства не означает отсутствия истины. Оно означает нечто гораздо более интересное: выбранный формальный язык и набор правил недостаточны для того, чтобы из собственных оснований добраться до этой истины.

Это различие необходимо удерживать особенно строго, потому что человеческий язык легко создаёт ложную интуицию. Мы привыкли говорить: «Если не доказано, значит, неизвестно, истинно ли это». Но «не доказано в системе» и «неизвестно вообще» — не одно и то же. Более того, «невозможно доказать в этой системе» и «невозможно узнать никаким способом» также не одно и то же. Можно перейти к более сильной системе аксиом, изменить формальный язык, добавить новые принципы и получить возможность доказать то, что раньше было недоказуемым. Однако вместе с усилением системы возникает новая граница. То, что было недостижимо на одном уровне, становится достижимым на другом, но абсолютного финального уровня, автоматически решающего все подобные вопросы, мы не получаем.

Чтобы понять глубину этой проблемы, необходимо отказаться от представления о математике как о едином гигантском хранилище уже существующих доказательств. Математика представляет собой скорее иерархию формальных структур, каждая из которых определяется своими исходными предпосылками, языком и правилами вывода. Когда математик говорит, что некоторое утверждение доказано, он фактически сообщает не только нечто об этом утверждении, но и нечто о системе, внутри которой было выполнено доказательство. Доказательство

всегда имеет контекст. Оно происходит не «в математике вообще», а относительно определённых исходных принципов.

Это обстоятельство кажется незначительной технической деталью до тех пор, пока мы не задаём вопрос о самой системе. Если формальная система умеет доказывать утверждения о числах, она в определённом смысле способна описывать собственные доказательства. Доказательство становится математическим объектом. Последовательность символов можно рассматривать как конечную структуру, проверять правильность её построения, кодировать рассуждения числами и говорить о свойствах этих закодированных последовательностей средствами самой арифметики. В этот момент граница между математическим объектом и описанием математического рассуждения начинает исчезать.

Именно здесь возникает принципиально новая ситуация. Формальная система, которая первоначально предназначалась для рассуждений о числах, получает возможность в определённой форме рассуждать о собственных доказательствах. Она начинает сталкиваться с самой собой. А как только система становится достаточно выразительной, чтобы представлять внутри себя некоторые утверждения о собственных процедурах доказательства, возникает возможность построить утверждение, содержание которого связано с его собственной доказуемостью.

В этом месте особенно важно не допустить популярного упрощения. Иногда теоремы Гёделя о неполноте пересказывают так, будто Гёдель просто «создал математический парадокс вроде утверждения „это утверждение недоказуемо“». Такое объяснение полезно только на первом уровне знакомства. На более серьёзном уровне оно скрывает главное. Гёделевская конструкция не является словесной игрой. Её сила заключается в том, что формальная система способна арифметически представить собственные операции с доказательствами. Самоссылка возникает не потому, что математик подставил хитрую фразу в текст, а потому, что достаточно выразительная система позволяет превратить метаматематическое рассуждение в объект самой математики.

Представим систему, которая содержит достаточно арифметики и в которой мы можем формально проверять, является ли заданная последовательность символов корректным доказательством. Для каждой такой последовательности в принципе можно определить, что она доказывает. Если система достаточно выразительна, утверждения о существовании доказательства могут быть переведены на язык арифметики. Иными словами, система способна говорить не только «существует такое число, которое обладает определённым свойством», но и в специальном смысле «существует код корректного доказательства данного утверждения».

Теперь появляется возможность построить утверждение, которое связывает собственную истинность с отсутствием доказательства внутри системы. Но здесь необходимо быть осторожным. В популярной форме это можно представить как утверждение: «Данное утверждение не имеет доказательства в этой системе». Если система непротиворечива и удовлетворяет необходимым условиям, то именно такая конструкция приводит к фундаментальному результату: система не может доказать соответствующее утверждение. В достаточно подходящей интерпретации оно оказывается истинным именно потому, что система не способна его доказать.

Возникает необычная картина. У нас имеется утверждение, которое не является бессмысленным. Оно не представляет собой ошибку записи, не нарушает правила языка и не является обычным логическим противоречием. Оно имеет вполне определённое математическое содержание. Однако система, внутри которой оно сформулировано, не может произвести доказательство этого содержания.

Это чрезвычайно важный момент. Неполнота не означает противоречивость. Противоречивая система — это система, в которой можно вывести несовместимые утверждения и тем самым разрушить надёжность формального вывода. Неполная система устроена иначе. Она может оставаться непротиворечивой и при этом не отвечать на некоторые вопросы, которые сама способна сформулировать.

Разница между противоречивостью и неполнотой фундаментальна. Если система противоречива, её проблема состоит в том, что правила позволяют получить слишком многое. Если система неполна, проблема противоположна: правила не позволяют получить всё, что система способна выразить. В одном случае система теряет надёжность из-за избытка выводимости, в другом — полноту из-за недостатка выводимости.

Для человеческой интуиции это выглядит почти парадоксально. Мы привыкли считать, что строгость системы должна в конечном итоге обеспечивать определённость. Чем точнее правила, тем меньше неопределённости. Но в достаточно богатой математике происходит неожиданное: увеличение строгости не устраняет всех неопределённых вопросов. Напротив, сама строгость позволяет настолько точно исследовать структуру формального рассуждения, что становится видна невозможность получить внутри данной системы некоторые ответы.

Здесь необходимо сделать ещё один шаг. Что означает слово «истинно» в утверждении о неполноте? Ведь если утверждение недоказуемо, на каком основании мы называем его истинным? Ответ требует различать внутреннюю и внешнюю перспективы.

Внутри формальной системы мы имеем правила, аксиомы и допустимые способы вывода. Система не может вывести определённое утверждение. Но математик, рассуждающий о системе извне, может анализировать её устройство и показать, что при условии непротиворечивости данное утверждение не может быть опровергнуто определённым образом и не может быть доказано самой системой. Если дополнительно известно, что система действительно непротиворечива, появляется основание говорить об истинности соответствующего утверждения в стандартной математической интерпретации.

Таким образом, возникает своеобразная лестница уровней. На первом уровне находится формальная система. На втором — математик, исследующий свойства этой системы. На третьем — возможна система, формализующая рассуждения о первой. Но если мы пытаемся включить рассуждающего наблюдателя в ту же систему, возникает новая проблема самоссылки. То, что на одном уровне кажется внешним метаязыком, на другом может стать объектом внутреннего анализа. И тогда возникает новая граница.

Это чрезвычайно важное следствие для понятия рациональности. Рациональность часто понимается как способность последовательно применять правила рассуждения и получать правильные выводы. Но если сама система правил имеет границы выразимости и доказуемости, то рациональность не может быть отождествлена с наличием алгоритма, который способен решить всякий сформулированный вопрос. Рациональный разум может столкнуться с вопросом, на который его собственная формальная архитектура не даёт ответа.

Более того, сам вопрос может быть поставлен на более высоком уровне. Если некоторый математический объект недоказуем в системе А, это ещё не означает, что он недоказуем вообще. Можно построить систему Б, которая содержит дополнительные аксиомы и способна доказать утверждение, недоступное системе А. Но затем теорема о неполноте применяется уже к системе Б. Если Б достаточно сильна и удовлетворяет необходимым условиям, у неё также возникают собственные недоказуемые утверждения.

Это напоминает лестницу, у которой нельзя просто добраться до последней ступени. Поднимаясь, мы действительно получаем больше возможностей. Сильная система способна доказать то, что слабая не могла доказать. Но усиление системы не уничтожает сам принцип существования границ. Оно переносит границу выше.

И здесь возникает одна из самых глубоких философских интерпретаций неполноты. Проблема заключается не в том, что человеку не хватило интеллекта, чтобы придумать правильное доказательство. Если утверждение действительно независимо от данной системы, то бессмысленно требовать от более быстрого или более умного математика доказательства, которое запрещено структурой самой системы. Можно представить интеллект, превосходящий человеческий в миллиарды раз, обладающий колоссальной памятью и практически неограниченными

вычислительными ресурсами. Если этот интеллект работает строго внутри той же формальной системы, то увеличение его мощности само по себе не устраняет её структурную неполноту.

Это различие между вычислительной мощностью и логической достижимостью является одним из центральных мотивов всей книги. Существуют задачи, которые трудно решить, потому что на них требуется много времени. Существуют задачи, которые практически невозможно решить из-за ограниченности ресурсов. Но существуют и задачи другого класса — такие, для которых определённый тип универсального решения невозможен принципиально. В последнем случае увеличение мощности не превращает невозможное в возможное.

Представим компьютер, который способен выполнять триллионы операций в секунду. Если некоторую задачу можно решить перебором, но для полного перебора требуется время, превосходящее возраст Вселенной, мы можем назвать её практически неразрешимой для данного компьютера. Однако в принципе задача остаётся вычислимой. Более мощная машина или более эффективный алгоритм могут изменить ситуацию. Неполнота имеет другую природу. Здесь проблема не обязательно в том, что доказательство слишком длинное. В некоторых случаях проблема в том, что никакого доказательства внутри данной системы не существует.

Различие между «очень трудно» и «невозможно в рамках системы» принципиально. Человеческий мозг постоянно сталкивается с задачами, которые слишком сложны для него практически, но не имеют никакого фундаментального запрета на решение. Теорема о неполноте указывает на другую категорию границ. Она заставляет нас спрашивать не только «сколько вычислений потребуется?», но и «существует ли вообще допустимая последовательность вычислений внутри выбранного формального пространства, ведущая к требуемому результату?».

Для интеллектуальной системы это различие особенно важно. Если мы создадим искусственный интеллект, способный перебрать больше вариантов, чем все люди вместе взятые, это ещё не означает, что он преодолеет логические границы формальной системы. Если же искусственный интеллект изменит саму систему — введёт новые аксиомы, расширит язык, изменит способ представления объектов, — он сможет выйти за прежнюю границу. Но тогда мы уже говорим не о преодолении неполноты вообще, а о переходе в более сильную систему.

Это можно выразить иначе: невозможно победить неполноту просто добавлением вычислительной мощности. Можно изменить пространство, в котором осуществляется рассуждение. Но изменение пространства порождает новые основания, а вместе с ними — новые границы.

Именно поэтому идея абсолютного математического разума оказывается значительно сложнее, чем идея очень мощного вычислителя. Пусть существует интеллект, способный мгновенно находить все доказательства в некоторой формальной системе. Даже тогда он не получает ответа на все математические вопросы. Он получает идеальный механизм поиска внутри определённого пространства доказательств. Но если существуют утверждения, которые не являются выводимыми в этом пространстве, скорость поиска не имеет значения.

Можно увеличить скорость до немыслимого значения. Можно представить вычислительную систему, работающую быстрее любого физически мыслимого компьютера. Можно вообразить интеллект, для которого человеческие математические доказательства выглядят как элементарные упражнения. Но если правила вывода остаются теми же, структурная граница остаётся на месте.

Это подводит нас к вопросу, который особенно важен для темы данной книги: означает ли неполнота ограниченность самого разума или только ограниченность формальных систем?

Здесь требуется осторожность. Теоремы Гёделя непосредственно говорят о свойствах определённых формальных систем. Нельзя без дополнительных аргументов утверждать, что человеческий разум буквально является такой системой и потому автоматически подчиняется теореме в полном объёме. Человеческое мышление не является просто набором аксиом и

механических правил вывода. Оно использует интуицию, аналогии, восприятие, язык, эксперименты, эвристики, изменение понятий и переход между различными системами описания.

Поэтому было бы ошибкой сказать: «Гёдель доказал, что человеческий мозг не способен познать определённые истины». Это слишком сильное утверждение. Гёдель показал другое: достаточно богатые формальные системы имеют принципиальные ограничения на доказуемость. Философский вопрос возникает тогда, когда мы пытаемся понять, что это означает для идеала полностью формализованного рационального интеллекта.

Если предположить, что весь человеческий математический разум можно полностью представить как некоторую фиксированную формальную процедуру, возникает серьёзная проблема. Для этой процедуры можно построить соответствующие утверждения, недоступные ей в качестве доказуемых. Но человек способен изменить свои методы, принять новые аксиомы, перейти к другой математической структуре. Следовательно, возможно, сила человеческого мышления состоит не в том, что оно является единой завершённой формальной системой, а в способности переходить между системами.

Однако такой переход не является бесплатным. Новая система требует новых оснований. Когда математик принимает дополнительную аксиому, он расширяет пространство доказуемого, но одновременно принимает новое исходное допущение. Он не получает истину из ничего. Он изменяет условия, в которых определяются допустимые доказательства.

Здесь возникает один из наиболее интересных вопросов философии математики: откуда берётся уверенность в новых аксиомах? Если прежняя система не могла решить некоторый вопрос, мы добавляем принцип, позволяющий его решить. Но почему именно этот принцип следует считать допустимым? Иногда ответ связан с математической плодотворностью, внутренней согласованностью, интуитивной убедительностью, связью с другими теориями или способностью систематизировать уже известные результаты. Но ни один из этих критериев не превращает новую аксиому в абсолютную истину, независимую от любого контекста.

Это означает, что математическое мышление способно расширяться, но его расширение имеет метаматематическую цену. Каждый новый уровень позволяет увидеть больше, но одновременно делает видимыми новые вопросы о самом этом уровне.

Именно здесь неполнота начинает приобретать философское значение. Она разрушает представление о математике как о потенциально конечной системе, которую можно однажды полностью закрыть. Если под «полной математикой» понимать единую достаточно выразительную и непротиворечивую формальную систему, способную решить каждое содержательное утверждение соответствующей области, то надежда на такую систему сталкивается с принципиальным препятствием.

Но особенно интересно другое: неполнота не уничтожает математику. Она не превращает математическую истину в субъективное мнение и не означает, что «всё относительно». Напротив, математика продолжает оставаться одной из наиболее строгих форм рационального знания. Теорема о неполноте показывает не беспомощность математического разума, а структуру его силы. Именно потому, что формальная система способна достаточно глубоко говорить о собственных доказательствах, мы можем обнаружить её ограничения.

Это напоминает ситуацию с картой. Чем более подробной становится карта, тем больше информации о территории она содержит. Но карта остаётся картой. Если попытаться сделать её абсолютно полной, возникает вопрос: должна ли она включать саму карту, лежащую на территории? А если включить её, нужно ли изображать ещё одну карту внутри первой? И так далее. Самоссылка порождает не просто проблему объёма, а проблему принципа построения описания.

В математике ситуация значительно строже, чем в этой аналогии. Формальная система способна кодировать собственные доказательства, и поэтому проблема самописания превращается из философской метафоры в точный математический результат. Система оказывается

способной формализовать утверждения о собственном поведении, а это открывает путь к построению утверждений, которые она сама не может решить.

Здесь особенно хорошо видно, почему истина не равна доказательству. Доказательство — это отношение между утверждением и конкретной системой правил. Истина — понятие, которое может рассматриваться относительно математической структуры или интерпретации. Эти понятия тесно связаны, но не совпадают.

Если некоторое утверждение доказано в системе, это означает, что существует допустимая последовательность шагов от принятых аксиом к данному утверждению. Но если доказательства нет, отсюда ещё не следует, что утверждение ложно. Оно может быть независимым от системы. Тогда система недостаточно сильна, чтобы решить вопрос.

Это особенно хорошо видно на примере геометрии. Если изменить исходные аксиомы, можно получить разные геометрические системы, каждая из которых обладает собственной внутренней логикой. Утверждение, обязательное в одной геометрии, может оказаться неверным в другой. Здесь становится очевидно, что математическая истина не существует в полном отрыве от структуры, относительно которой формулируется утверждение.

Но неполнота идёт дальше. Даже после выбора достаточно конкретной формальной системы остаются утверждения, которые нельзя решить средствами этой системы. И если мы расширим её, новая система также может оказаться неполной. Поэтому проблема не сводится к неправильному выбору аксиом. Она касается самого отношения между выразимостью и доказуемостью.

Можно представить это как бесконечное движение границы. На каждом уровне у нас есть пространство утверждений, которые можно доказать. За его пределами находятся утверждения, которые система не способна решить. Мы добавляем новые принципы и расширяем пространство. Некоторые прежние вопросы получают ответы. Но вместе с расширением возникает новый внешний край.

Это не означает, что все уровни одинаково ценны или что невозможно строить всё более сильные математические теории. Напротив, история математики показывает противоположное. Новые аксиомы, методы и концепции способны радикально расширять область доказуемого. Но ни один конкретный уровень не получает автоматически статус окончательного языка всей математической истины.

Для идеи сверхразума это имеет особое значение. Допустим, мы создаём интеллект, который во много раз превосходит человека. Он способен анализировать математические структуры, создавать новые доказательства, находить скрытые закономерности и конструировать новые формальные системы. Возникает соблазн сказать: если интеллект достаточно мощен, он когда-нибудь сможет решить всё.

Но «решить всё» здесь необходимо определить. Если имеется в виду решить все задачи в некоторой фиксированной системе, то даже очень мощный интеллект сталкивается с её внутренними пределами. Если имеется в виду переходить от системы к системе и расширять область доказуемого, тогда можно двигаться дальше. Но уже нельзя говорить о завершённом абсолютном алгоритме познания. Каждый переход открывает новое пространство вопросов.

Сверхразум в таком случае может оказаться не существом, которое знает все ответы, а системой, способной эффективнее человека обнаруживать границы ответов. Его преимущество будет заключаться не только в способности доказать больше теорем, но и в способности быстрее определить, в какой системе находится вопрос, какие предпосылки необходимы, где проходит граница доказуемости и какие новые математические конструкции способны эту границу переместить.

Это гораздо более глубокое понимание интеллекта, чем простое увеличение скорости вычислений. Интеллектуальная мощность может выражаться в способности выбирать подхо-

дящую систему координат для самого вопроса. Иногда невозможно получить ответ, не изменив язык, в котором задан вопрос.

Здесь мы возвращаемся к теме предыдущих глав. Вопрос может быть истинным, но недоказуемым в определённой системе. Он может быть доказуемым после расширения системы. Он может зависеть от выбранных аксиом. Он может быть сформулирован так, что сама его постановка предполагает определённую концептуальную рамку. Поэтому зрелое математическое мышление должно исследовать не только ответ, но и пространство, в котором этот ответ ищется.

Возникает ещё более радикальный вопрос: может ли существовать разум, который способен одновременно видеть все такие уровни и выбрать окончательную систему, включающую все возможные истины?

Проблема состоит в том, что слово «окончательная» здесь становится подозрительным. Если система достаточно богата, чтобы выражать арифметику, и удовлетворяет соответствующим условиям, для неё возникают новые недоказуемые утверждения. Если мы создаём более сильную систему, она получает новые возможности, но и для неё возникает аналогичная граница. Поэтому сама идея окончательной формальной системы требует очень осторожного определения.

Можно сказать, что разум способен не только решать задачи, но и создавать новые математические уровни. Однако способность переходить к следующему уровню не означает достижение последнего. Возможно, сама структура математического познания такова, что процесс расширения принципиально открыт.

Это имеет прямое отношение к понятию абсолютного знания. Если под абсолютным знанием понимать обладание одним завершённым формальным механизмом, который способен доказать или опровергнуть всякое математически содержательное утверждение, то неполнота показывает серьёзность проблемы. Если же под абсолютным знанием понимать способность потенциально переходить к новым уровням формализации, то это уже не конечное состояние знания, а бесконечный процесс его расширения.

Разница между этими двумя представлениями огромна. В первом случае знание — это гигантская, но конечная библиотека, в которой когда-нибудь можно собрать все ответы. Во втором — это иерархия всё более мощных способов постановки и решения вопросов. Тогда граница знания не является стеной, которую можно однажды разрушить. Она напоминает горизонт, который перемещается вместе с наблюдателем.

Но здесь возникает опасность философского злоупотребления теоремой Гёделя. Нельзя использовать неполноту как универсальное объяснение любой тайны. Неполнота не доказывает существование сверхъестественного. Она не доказывает невозможность познания сознания. Она не доказывает превосходство человеческой интуиции над вычислением. Она не утверждает, что всякая человеческая мысль находится за пределами формальной логики. И тем более она не означает, что любое непонятное утверждение автоматически является «истиной за пределами науки».

Научная строгость здесь особенно важна. Теорема о неполноте имеет точные условия применимости. Её сила именно в том, что она говорит меньше, чем ей иногда приписывают, но то, что она говорит, чрезвычайно глубоко. Она показывает, что определённый класс достаточно выразительных формальных систем не может одновременно обладать всеми желаемыми свойствами полноты и непротиворечивости.

Философский вывод должен строиться поверх этого точного результата, а не подменять его красивой метафорой. Самое интересное заключается в том, что даже строгое ограничение доказуемости заставляет пересмотреть идею рациональности. Рациональность больше не выглядит как способность получить ответ на любой вопрос. Она становится способностью понимать, какие ответы следуют из каких оснований, какие вопросы находятся внутри данной

системы, какие требуют её расширения, а какие вообще не имеют смысла без дополнительной структуры.

Это уже другая концепция интеллекта. Слабый интеллект спрашивает: «Как мне получить ответ?» Более сильный спрашивает: «Какие методы позволяют получить ответ?» Ещё более сильный спрашивает: «Почему этот вопрос должен иметь ответ внутри выбранной системы?» И, наконец, интеллектуально зрелая система способна спросить: «Какие ограничения самой системы определяют то, что я считаю допустимым ответом?»

На этом уровне неполнота перестаёт быть исключительно результатом математической логики. Она становится уроком о природе рационального мышления вообще. Любая интеллектуальная система действует в некотором пространстве представлений. Внутри этого пространства она может быть чрезвычайно мощной. Но мощность внутри пространства не гарантирует способности выйти за пределы самого пространства.

Человек способен изменить свои математические основания. Он способен создать новую теорию, новый язык, новый тип объекта. В этом смысле человеческий разум не заключён навсегда в одной формальной системе. Но каждый раз, когда он делает такой переход, он создаёт новую систему, а вместе с ней — новые вопросы о её границах.

Именно поэтому история математики не похожа на постепенное заполнение заранее существующей энциклопедии. Она скорее напоминает расширение самого пространства мысли. Появление новых чисел, геометрий, алгебраических структур, логик, теорий множеств и вычислительных моделей не просто добавляло новые ответы. Оно изменяло то, какие вопросы вообще можно было сформулировать.

В этом смысле неполнота имеет почти парадоксальное значение. Она не говорит, что математический разум слишком слаб. Она показывает, что достаточно сильный разум способен увидеть собственную границу. Система должна быть достаточно выразительной, чтобы поставить вопрос о собственных доказательствах, а затем оказывается не способна полностью закрыть этот вопрос собственными средствами.

Это одна из наиболее необычных черт математического знания: развитие разума может увеличивать не только область известного, но и точность знания о неизвестном. В примитивной картине незнание выглядит как пустота. В зрелой математике незнание может иметь структуру. Мы можем знать, что определённое утверждение недоказуемо в данной системе. Мы можем знать, какие дополнительные предпосылки сделали бы его доказуемым. Мы можем знать, что усиление системы породит новые независимые вопросы. Незнание перестаёт быть простой дырой в знаниях и становится математически описываемым объектом.

И здесь появляется, возможно, наиболее важный философский вывод этой главы. Истина и доказательство связаны, но не тождественны. Доказательство — это путь, разрешённый определённой системой. Истина — это свойство утверждения относительно определённой математической структуры или интерпретации. Когда система сталкивается с утверждением, которое она не может доказать, это не обязательно означает поражение истины. Иногда это означает, что сама система достигла границы собственных средств выражения и обоснования.

Разум, который не понимает этой разницы, будет принимать собственную доказательную систему за саму реальность. Он будет считать истинным только то, что умеет вывести, и ложным всё остальное. Такой разум может быть безупречно последовательным и одновременно интеллектуально ограниченным.

Разум более высокого уровня способен сделать следующий шаг. Он может исследовать не только утверждения внутри системы, но и свойства самой системы. Он может сравнивать разные основания, понимать зависимость выводов от аксиом, выявлять независимые утверждения и строить новые формальные пространства. Но и здесь нет гарантии окончательной завершённости. Каждый новый уровень потенциально становится объектом следующего метауровня.

Поэтому вопрос «что способен понять самый умный разум?» приобретает совершенно иной смысл. Возможно, предел интеллекта определяется не количеством информации и не скоростью вычислений. Возможно, один из фундаментальных пределов возникает там, где разум пытается сделать собственную систему рассуждения полностью прозрачной для самой себя.

На этом месте мы подходим к следующему вопросу. Неполнота показывает, что внутри достаточно сильной формальной системы существуют утверждения, которые она не способна решить. Но можно поставить ещё более конкретную и жёсткую задачу: существует ли универсальный способ заранее определить, какие вычислительные процессы когда-нибудь завершатся, а какие будут продолжаться бесконечно?

Если бы такой способ существовал, значительная часть неопределённости вычислимого мира могла бы быть устранена одним универсальным анализатором. Любую программу можно было бы передать ему, и он заранее сообщил бы, завершится ли её работа. Но именно здесь возникает следующий фундаментальный барьер математического разума — **проблема остановки**.

В следующей главе мы увидим, что невозможность универсального предсказателя вычислительного поведения возникает не из-за недостатка мощности компьютеров и не из-за слишком большого объёма вычислений. Она возникает из самой логической структуры вычисления. И это позволит сделать ещё один шаг в исследовании вопроса, который проходит через всю эту книгу: где заканчивается трудность задачи и начинается принципиальная невозможность её решения?

Глава 7. Невычислимое

До сих пор мы рассматривали пределы математического разума главным образом через понятие доказательства. Формальная система может быть достаточно строгой, достаточно выразительной и достаточно мощной, но при этом сталкиваться с утверждениями, которые невозможно доказать внутри неё. Теперь необходимо сделать следующий шаг. Даже если отвлечься от вопроса математической истины и обратиться к самой процедуре вычисления, обнаруживается граница, которая выглядит ещё более механически и потому особенно впечатляет: существуют задачи, для которых не существует универсального алгоритма решения.

Это утверждение требует предельной точности. Оно не означает, что некоторые задачи просто слишком сложны для современных компьютеров. Не означает оно и того, что вычисление может занять миллион, миллиард или даже невероятно большое количество лет. Во всех этих случаях мы всё ещё имеем дело с вычислимой задачей. Если существует алгоритм, который в принципе выдаёт правильный ответ за конечное число шагов, задача относится к вычислимым, даже если практически выполнить этот алгоритм невозможно.

Невычислимость начинается в другом месте. Здесь речь идёт о задачах, для которых невозможно построить один универсальный алгоритм, способный получать правильный ответ для всех допустимых входных данных. Можно представить сколь угодно мощную вычислительную машину, практически неограниченную память и фантастическую скорость обработки информации. Если требуемый универсальный алгоритм не существует, увеличение вычислительной мощности не создаёт его.

Это различие между трудностью и невозможностью является одним из центральных различий всей теории вычислений. Человек часто воспринимает компьютер как машину, которая отличается от человеческого мозга прежде всего скоростью. Если задача слишком трудна для человека, можно передать её компьютеру. Если недостаточно одного компьютера, можно использовать тысячу. Если недостаточно тысячи, можно построить вычислительный кластер. Если недостаточно кластера, можно увеличить память, число процессоров и скорость операций. Такая логика действительно работает для огромного количества практически сложных задач.

Но она не работает для принципиально невычислимых задач.

Чтобы понять почему, необходимо сначала освободиться от слишком узкого представления о вычислении. Вычисление — это не обязательно современный компьютер, экран, процессор и программа. В более фундаментальном смысле вычисление представляет собой механическую процедуру преобразования заданной информации по определённым правилам. Если существует конечная последовательность однозначных операций, которая для каждого допустимого входа позволяет получить требуемый результат, перед нами вычислимая задача.

Компьютер является лишь физической реализацией таких процедур. Его процессор, память, язык программирования и операционная система — конкретные уровни реализации. На более абстрактном уровне принцип один и тот же: имеется исходное описание, имеются правила обработки, и в результате получается новое описание.

Эта абстракция чрезвычайно важна. Она позволяет отделить свойства вычисления от свойств конкретного устройства. Если задача решается одним универсальным компьютером, она в соответствующем смысле может быть решена и другим универсальным компьютером, если у него достаточно ресурсов и времени. Поэтому увеличение мощности конкретной машины не меняет фундаментального класса задачи.

Можно провести простое различие между тремя ситуациями.

В первой ситуации задача вычислима и практически проста. Компьютер решает её за доли секунды.

Во второй ситуации задача вычислима, но практически чудовищно сложна. Теоретически алгоритм существует, однако количество необходимых операций настолько велико, что реальное выполнение оказывается недостижимым.

В третьей ситуации универсального алгоритма вообще не существует. Здесь бессмысленно спрашивать, насколько мощным должен быть компьютер. Проблема находится не на уровне оборудования. Не хватает не мощности, а самого алгоритма.

Именно третья категория представляет для нашего исследования особый интерес.

Первые две ситуации можно сравнить с путешествием. В одном случае пункт назначения находится рядом. В другом — чрезвычайно далеко, и дорога занимает миллионы лет. Но дорога существует. Невычислимость — это ситуация иного рода: проблема не в длине дороги, а в том, что универсальной дороги данного типа не существует.

Один из наиболее знаменитых примеров связан с программами и их поведением. Рассмотрим любую программу. Она получает входные данные и начинает выполнять инструкции. Иногда она заканчивает работу и выдаёт результат. Иногда она продолжает работать бесконечно. Иногда она попадает в бесконечный цикл, иногда постоянно порождает новые вычисления, иногда ждёт событие, которое никогда не произойдёт.

Возникает совершенно естественный вопрос: можно ли создать универсальный анализатор программ, которому мы передадим описание любой программы и её входных данных, а он заранее определит, завершится ли выполнение?

На первый взгляд кажется, что такая задача должна быть решаемой. Ведь программа состоит из конечного количества инструкций. Её поведение определяется правилами. Если мы достаточно хорошо проанализируем эти правила, почему бы не определить результат заранее?

Можно даже представить чрезвычайно мощную систему анализа. Она не запускает программу обычным способом, а исследует её структуру. Она умеет строить модели выполнения, отслеживать все возможные переходы, распознавать циклы, анализировать зависимости переменных и находить скрытые закономерности. Она работает намного умнее обычного компьютера.

Но универсального анализатора такого типа не существует.

Причина этого результата гораздо глубже, чем может показаться. Проблема не в том, что программы иногда становятся слишком длинными. Не в том, что количество возможных состояний превышает объём памяти. И не в том, что человечество ещё не придумало достаточно хитрый метод анализа.

Проблема возникает из-за самоссылки.

Представим, что универсальный анализатор всё-таки существует. Он получает программу и отвечает на вопрос, завершится ли она. Теперь можно создать специальную программу, которая получает описание некоторой программы и обращается к универсальному анализатору с вопросом о её поведении. Если анализатор утверждает, что проверяемая программа завершится, специальная программа намеренно продолжает работать бесконечно. Если анализатор утверждает, что проверяемая программа будет работать бесконечно, специальная программа, наоборот, завершает работу.

Теперь применим универсальный анализатор к этой специальной программе, которая анализирует собственное поведение.

Если анализатор скажет, что она завершится, она сделает противоположное и не завершится.

Если анализатор скажет, что она не завершится, она сделает противоположное и завершится.

Таким образом, предполагаемый универсальный предсказатель неизбежно ошибается.

Это не практическая демонстрация того, что компьютеру не хватает мощности. Это логическое доказательство того, что универсальный алгоритм подобного типа невозможен.

Именно здесь теория вычислимости совершает важнейший переход от инженерии к логике. До этого момента можно было думать, что достаточно мощная машина в конечном счёте сможет проанализировать любую программу. Теперь становится ясно: существует класс вопросов о вычислениях, для которых не существует общего алгоритмического решения.

Особенно важно понять, что отсутствие универсального решения не означает отсутствия отдельных решений. Для конкретной программы иногда можно совершенно точно определить, завершится она или нет. Более того, многие практически используемые программы можно анализировать достаточно хорошо. Существуют статические анализаторы, системы проверки программ, методы доказательства корректности, средства обнаружения бесконечных циклов и множество других инструментов.

Невычислимость появляется именно тогда, когда мы требуем универсальности.

Это принципиальная граница. Можно создать алгоритм, который прекрасно работает на огромном классе программ. Можно добавить новые эвристики. Можно совершенствовать анализатор. Можно использовать машинное обучение, символьные методы, формальную верификацию, огромные вычислительные мощности. Но нельзя получить единственный универсальный алгоритм, который для любой возможной программы и любого входа безошибочно решает вопрос о её остановке.

Это означает, что существуют программы, поведение которых невозможно заранее классифицировать посредством общего механического метода.

Здесь появляется ещё одна важная мысль. Невычислимость не означает непредсказуемость в бытовом смысле. Программа может быть полностью детерминированной. В ней может не быть никакой случайности. Каждая инструкция однозначно определяет следующую. И всё же не существует универсальной процедуры, которая заранее решает для всех таких программ, чем закончится их выполнение.

Это один из самых неожиданных результатов теории вычислений. Детерминизм не равен предсказуемости.

Если мы знаем правила физической системы и её начальное состояние, нам может казаться, что будущее должно быть принципиально вычислимо. Но знание правил и наличие универсального алгоритма предсказания — разные вещи. Динамика может быть полностью определена, а общий способ заранее получить требуемый ответ всё равно может отсутствовать.

Разница между этими понятиями имеет фундаментальное значение для разговора об интеллекте. Интеллект часто описывают как способность предсказывать. Чем умнее система, тем дальше она видит последствия своих действий. Но из теории вычислимости следует, что предсказательная сила не может быть универсальной даже в мире чисто формальных вычислений.

Это особенно интересно потому, что вычислительные системы способны моделировать самих себя. Программа может анализировать другие программы. Компьютер может запускать симуляцию компьютера. Искусственный интеллект может анализировать алгоритм, который используется для анализа искусственного интеллекта. Но как только система получает достаточно выразительные средства для описания собственного поведения, возникает возможность самоссылочной конструкции.

Именно самоссылка связывает эту главу с предыдущей. Неполнота и невычислимость не являются двумя совершенно изолированными загадками. Они возникают из близкого фундаментального источника: достаточно выразительная система получает возможность в определённом смысле описывать собственные процессы, после чего возникает принципиальное ограничение на её способность полностью контролировать или предсказывать их.

Но между этими результатами существует важное различие. Теоремы о неполноте относятся к доказуемости утверждений внутри формальных систем. Теория вычислимости рассматривает возможность существования алгоритмов, решающих определённые классы задач.

В одном случае возникает вопрос «можно ли доказать?», в другом — «существует ли процедура, которая всегда выдаёт правильный ответ?».

Эти вопросы тесно связаны, но не совпадают.

Можно представить себе формальную систему как пространство доказательств, а вычислительную систему — как пространство процедур. В обоих случаях возникает граница, которую нельзя устранить простым увеличением ресурсов.

Теперь необходимо обратиться к распространённому заблуждению: если универсальный алгоритм невозможен, может быть, его заменит достаточно развитый искусственный интеллект.

Ответ зависит от того, что именно понимать под искусственным интеллектом.

Если искусственный интеллект является вычислительной системой, работающей посредством алгоритмов, то сама по себе его интеллектуальная мощность не уничтожает невычислимость. Он может быть невероятно сложным. Он может использовать миллиарды взаимосвязанных процессов. Он может обучаться. Он может создавать новые программы. Он может исследовать математические доказательства, строить гипотезы и находить неожиданные решения. Но если требуемая задача относится к классу принципиально невычислимых, то не существует алгоритмического механизма, который превратил бы её в универсально решаемую только за счёт увеличения сложности системы.

Это не означает, что искусственный интеллект ничего не сможет сделать. Напротив, он может решать конкретные экземпляры задач, для которых не существует универсального решения. Он может использовать дополнительные знания о конкретной ситуации. Может ограничить класс программ. Может доказать остановку конкретного алгоритма. Может обнаружить закономерность там, где простой анализатор бессилен.

Но как только мы возвращаемся к требованию «для любой возможной программы», граница появляется снова.

Именно требование универсальности превращает многие задачи в принципиально неразрешимые.

Это имеет огромное значение для проектирования интеллектуальных систем. Инженер может создать анализатор, который работает почти идеально на практическом наборе программ. Затем появляется соблазн сказать, что проблема решена. Но теоретическая граница никуда не исчезает. Просто реальный класс задач оказался намного уже, чем класс всех возможных программ.

Так работает значительная часть прикладной информатики. Формально невозможная универсальная задача может иметь чрезвычайно эффективное практическое приближение в ограниченной области. Мы можем проверять безопасность программ, находить ошибки, доказывать свойства алгоритмов, анализировать зависимости и обнаруживать множество классов проблем. Невычислимость не запрещает полезность. Она запрещает универсальность.

Это различие чрезвычайно важно и за пределами информатики.

Представим медицинскую систему, которая пытается предсказывать течение болезни. Для ограниченного класса пациентов она может давать исключительно точные прогнозы. Но из этого не следует существование универсального алгоритма, способного предсказать любую индивидуальную траекторию любого организма при любых обстоятельствах.

Представим экономическую систему, которая прогнозирует поведение рынка. Даже если она анализирует огромные объёмы информации, это не означает существования универсального предсказателя, который заранее определяет любое будущее состояние экономики.

Представим систему, моделирующую человеческое поведение. Она может успешно предсказывать множество решений конкретного человека, но отсюда не следует, что существует универсальная процедура, способная вычислить все будущие действия любого возможного человека при любых условиях.

Однако здесь необходимо избегать чрезмерных аналогий. Невычислимость конкретной математической задачи не доказывает автоматически невычислимость любой физической, биологической или социальной системы. Чтобы переносить выводы теории вычислимости на реальный мир, необходимо показать, что соответствующая система действительно обладает необходимой структурой и что рассматриваемая задача формализуется соответствующим образом.

Тем не менее общий философский урок остаётся чрезвычайно сильным: детальное знание правил не гарантирует существования универсального алгоритма, который извлекает из этих правил любой интересующий нас прогноз.

Это разрушает очень глубокую разновидность механистического оптимизма. Согласно ему, если мир достаточно хорошо описан и вычислительная мощность достаточно велика, любое неизвестное в конечном счёте превращается в вычислимое. Теория вычислимости показывает, что такое превращение не является универсальным законом.

Есть вопросы, которые можно поставить даже внутри полностью формализованного мира, но для которых не существует общего алгоритмического ответа.

Причём это не обязательно означает, что конкретный наблюдатель никогда не сможет получить ответ. Это означает, что невозможно построить одну универсальную механическую процедуру, гарантированно отвечающую на все случаи.

Такое различие кажется небольшим, но оно фундаментально.

Если спросить, завершится ли конкретная программа, ответ иногда можно получить простым анализом. Если программа состоит из нескольких инструкций и не содержит циклов, вопрос тривиален. Если она имеет очевидный бесконечный цикл, ответ также может быть очевиден. Но когда мы требуем метод, который одинаково надёжно работает для всех программ, задача меняет свой характер.

Это напоминает различие между существованием решения и существованием метода получения решения в каждом случае.

В математике подобные различия встречаются постоянно. Можно доказать существование объекта, не имея простого способа его построить. Можно знать, что определённое множество обладает свойством, не располагая эффективным способом вычислить соответствующий элемент. Можно показать, что решение существует, но не получить алгоритма, который универсально его находит.

Вычислимость требует не просто существования ответа. Она требует существования эффективной процедуры, которая в соответствии с точным алгоритмическим правилом получает ответ.

И здесь появляется ещё одна важная граница: между математическим существованием и конструктивной доступностью.

Человек способен рассуждать о математических объектах, которые нельзя эффективно вычислить. Можно определить свойства объекта, доказать его существование, исследовать его отношения с другими объектами — и при этом не иметь алгоритма, который для произвольного случая выдаёт требуемое конкретное значение.

Это снова возвращает нас к различию между истиной и процедурой. Истинное утверждение может существовать без доступного алгоритма его получения.

Следовательно, идея «разум знает всё, что истинно» становится ещё более проблематичной. Даже если отказаться от физических ограничений памяти и скорости, остаётся вопрос о том, какие истины вообще являются алгоритмически доступными.

Здесь необходимо ввести ещё одну тонкость. Невычислимость не является синонимом случайности.

Случайный процесс может не иметь заранее определённого результата. Невычислимый объект или процесс может быть полностью определён. Его поведение может быть однозначно

задано правилами. Ограничение возникает не в объекте, а в возможности построить универсальную процедуру, извлекающую из его описания требуемую информацию.

Это различие особенно важно для понимания цифрового мира. Компьютерная программа кажется полностью прозрачной, потому что она создана человеком. Но из конечного текста программы может следовать поведение, которое невозможно универсально предсказать алгоритмическим способом.

Получается почти парадоксальная ситуация: мы способны полностью записать правила поведения системы, но не способны создать универсальный алгоритм, который заранее определит все следствия этих правил.

Описание системы и полное предсказание системы — не одно и то же.

Это чрезвычайно глубокий принцип. Знать закон не означает уметь эффективно вычислить любое следствие этого закона. Иметь компактное описание системы не означает иметь компактный или вообще существующий алгоритм получения любого интересующего нас свойства.

В физике эта идея имеет особую осторожную интерпретацию. Мы не должны автоматически объявлять физическую Вселенную невычислимой только потому, что существуют невычислимые математические функции. Физические законы могут обладать дополнительной структурой, которая ограничивает реализуемые процессы. Но сама возможность невычислимости показывает, что переход от «у нас есть точное описание правил» к «мы можем алгоритмически предсказать всё» логически необоснован.

И здесь возникает вопрос о самой Вселенной как вычислительной системе. Если представить Вселенную как гигантский процесс обработки информации, возникает соблазн сказать: достаточно мощный компьютер мог бы смоделировать её целиком. Но что значит «достаточно мощный»? Если симуляция должна сохранять всё состояние системы с полной точностью, то симулятор сам становится физической системой. Тогда возникают ограничения памяти, энергии, времени и пространства. А если требуется моделировать собственное устройство, появляется проблема самоссылки.

Мы снова сталкиваемся с мотивом, который проходил через предыдущие главы: система, пытающаяся построить полную модель самой себя, оказывается в особом положении.

Вычислительная теория добавляет к этому ещё одну границу. Даже если мы не требуем моделировать систему целиком, некоторые вопросы о вычислительных процессах не имеют универсального алгоритмического решения.

Это означает, что идеал «абсолютного предсказателя» сталкивается сразу с несколькими независимыми препятствиями.

Первое — физическое: ограничены время, энергия, память и доступная информация.

Второе — информационное: часть состояния системы может быть недоступна наблюдателю.

Третье — логическое: некоторые выводы не следуют из выбранных оснований.

Четвёртое — вычислительное: некоторые классы задач не имеют универсальных алгоритмов решения.

Именно совокупность этих границ делает понятие абсолютного интеллекта значительно более проблематичным, чем просто «очень умный компьютер».

Но здесь можно пойти ещё дальше. Если искусственный интеллект способен создавать программы, анализировать программы и создавать новые версии самого себя, то проблема вычислимости начинает непосредственно касаться самосовершенствования.

Представим систему, которая пытается определить, улучшит ли некоторый предполагаемый модуль её собственную производительность. Она анализирует новый алгоритм, моделирует его поведение и пытается заранее решить, станет ли новая версия интеллектуально сильнее старой.

Для отдельных случаев это возможно. Но универсальный вопрос «улучшит ли любая возможная программа данную систему?» может быть настолько сложен, что его полное автоматическое решение сталкивается с фундаментальными ограничениями.

Это имеет значение для идеи рекурсивного самосовершенствования. Нельзя предполагать, что достаточно умная система обязательно сможет заранее вычислить все последствия собственной модификации. Система может создать новую версию, превосходящую её по некоторым параметрам, но не обладать полной формальной гарантией того, к чему приведёт каждая возможная модификация.

Таким образом, способность создавать более умные системы не равна способности полностью предсказать созданные системы.

Это различие становится особенно важным, когда речь идёт о сверхинтеллекте. Человек может создать программу, поведение которой не способен полностью понять. Более интеллектуальная система может создать другую систему, поведение которой уже она сама не способна универсально предсказать. Развитие интеллекта не обязано приводить к исчезновению границ предсказуемости.

Возможна даже ситуация, в которой система создаёт новый интеллектуальный уровень, способный решать задачи, недоступные создателю, но при этом создатель не имеет полного описания того, как именно этот интеллект будет приходить к решениям.

Это не обязательно означает хаос. Это может быть совершенно детерминированный процесс.

Здесь возникает фундаментальная асимметрия: способность породить вычислительный процесс может оказаться сильнее способности универсально предсказать его.

Мы уже встречались с похожей асимметрией в предыдущей главе. Формальная система способна выразить утверждение, которое не может доказать. Теперь вычислительная система может породить поведение, которое не может быть универсально предсказано соответствующим алгоритмом.

Оба результата показывают одну и ту же общую тенденцию: достаточно выразительная система может содержать больше структурного содержания, чем способен полностью охватить один фиксированный механизм анализа.

Это очень важная мысль для понимания интеллекта. Интеллект не следует представлять как универсальный ключ, который подходит ко всем замкам. Можно создать интеллект, способный строить всё более мощные модели, но это не гарантирует существования единой процедуры, которая автоматически решает любой вопрос, возникающий в пределах всех возможных моделей.

Возможно, одна из главных функций высокого интеллекта состоит не в устранении всех границ, а в различении их типов.

Одна задача требует больше памяти.

Другая требует более эффективного алгоритма.

Третья требует новой математической модели.

Четвёртая зависит от неизвестных данных.

Пятая неразрешима в выбранной формальной системе.

Шестая не имеет универсального алгоритмического решения.

Если интеллектуальная система не различает эти случаи, она будет совершать фундаментальную ошибку. Она будет относить всякую невозможность к недостатку вычислительных ресурсов. Тогда любое ограничение будет восприниматься как временное: «нужно просто больше памяти», «нужно больше данных», «нужен более быстрый процессор», «нужно обучить более крупную модель».

Но некоторые границы не исчезают после увеличения ресурсов.

Представим компьютер, который выполняет миллиард операций за секунду. Удвоим его мощность. Затем увеличим её в тысячу раз. Затем представим машину, которая выполняет операции практически мгновенно. Всё это может радикально изменить время решения огромного количества задач. Но никакая скорость не создаёт алгоритм, которого не существует.

Это особенно легко забыть в эпоху стремительного развития вычислительной техники. Практический опыт последних десятилетий формирует впечатление, что почти любое ограничение является временным. Недостаточно памяти — память увеличивается. Не хватает скорости — появляются новые процессоры. Недостаточно данных — создаются новые системы сбора информации. Не хватает алгоритма — разрабатывается новый метод.

История вычислений действительно показывает грандиозную силу такого подхода. Но теория вычислимости напоминает: этот процесс не бесконечен в универсальном смысле.

Существуют границы, которые невозможно преодолеть улучшением того же самого типа вычислительной машины.

Чтобы выйти за них, необходимо изменить постановку задачи, ограничить класс объектов, принять дополнительные знания или отказаться от требования универсальности.

И это чрезвычайно важное отличие. Когда инженер говорит: «Наш алгоритм не умеет решать эту задачу», это может быть недостатком конкретного алгоритма. Когда теория говорит: «Не существует алгоритма, который решает все задачи этого класса», речь идёт уже о другом уровне знания.

В первом случае мы ищем лучший метод.

Во втором случае мы исследуем границы самого понятия метода.

Именно поэтому теория вычислимости является не просто разделом информатики. Она является своеобразной философией алгоритмического разума. Она отвечает не только на вопрос «что умеют компьютеры?», но и на вопрос «что вообще может быть достигнуто посредством алгоритмической процедуры?».

Этот вопрос непосредственно связан с понятием рациональности. Современный рационализм часто предполагает, что если задача сформулирована достаточно точно, то рациональный агент должен искать алгоритм её решения. Но теория вычислимости показывает, что точная постановка вопроса ещё не гарантирует существования процедуры, которая его решает.

Иногда наиболее рациональным результатом является не решение, а доказательство отсутствия универсального решения.

Это принципиально меняет смысл слова «неудача». Если алгоритм не смог решить конкретную задачу, возможно, он плох. Если доказано, что универсального алгоритма не существует, невозможность становится знанием.

Так появляется ещё одна важная форма интеллектуальной силы: способность доказать невозможность определённого метода.

Слабый интеллект ищет ответ.

Более сильный интеллект ищет способ получить ответ.

Ещё более сильный интеллект способен исследовать, существует ли вообще такой способ.

Именно на этом уровне теория вычислимости становится частью философии интеллекта.

Однако здесь возникает новая трудность. Если универсального алгоритма не существует, можно ли заменить алгоритм интеллектуальной интуицией? Может ли человек или сверхразум делать то, что никакая формальная процедура не способна сделать?

На этот вопрос нельзя дать простой ответ.

Человек действительно способен использовать эвристики, догадки, аналогии и творческие переходы, которые не выглядят как заранее заданный алгоритм. Математик может увидеть структуру там, где механический перебор не помогает. Учёный может предложить совершенно новый способ постановки задачи. Исследователь может отказаться от первоначальной модели и создать другую.

Но из этого не следует автоматически, что человеческий разум обладает универсальным средством решения невычислимых задач. Чтобы утверждать такое, пришлось бы точно определить процедуру, посредством которой разум получает соответствующие ответы, и показать, что она действительно универсальна.

Пока этого нет, мы должны различать творческое мышление и магическую способность выходить за пределы вычислимости.

Человеческая мысль может менять алгоритмы. Она может создавать новые языки. Она может выбирать другие основания. Она может вводить дополнительные предположения. Но если результат можно систематически воспроизводить в форме определённой процедуры, возникает новый алгоритм, а не чудесный выход за пределы вычислимости.

Это возвращает нас к одной из центральных идей всей книги: интеллектуальная мощность может проявляться не в нарушении ограничений, а в способности менять уровень описания.

Если задача нерешаема в исходной системе, можно изменить систему.

Если невозможно предсказать программу универсальным анализатором, можно ограничить класс программ.

Если невозможно получить ответ из имеющихся данных, можно изменить эксперимент.

Если доказательство невозможно из существующих аксиом, можно исследовать новые основания.

Но каждый такой переход имеет цену. Мы не уничтожаем границу. Мы перестраиваем пространство, в котором она находится.

Именно поэтому слово «невычислимое» не следует понимать как обозначение какого-то таинственного мира за пределами математики. Невычислимое — это вполне строгая математическая категория. Она относится к задачам, для которых не существует соответствующей универсальной вычислительной процедуры.

В этом смысле невычислимое не является противоположностью разумному. Наоборот, обнаружение невычислимости является одним из величайших достижений рационального мышления. Разум способен доказать существование границы, которую никакой алгоритм данного типа не сможет пересечь.

Это особенно впечатляет потому, что доказательство невозможности само является результатом вычислимого и логического рассуждения. Мы используем строгую математику, чтобы показать границу того, что строгая алгоритмическая процедура может сделать.

Разум исследует собственный предел средствами самого разума.

В этом есть глубокий парадокс. Система оказывается способной доказать, что существуют задачи, которые она не может решить универсальным способом. Она не становится от этого всесильной. Но она получает знание о собственной несостоятельности в определённой области.

Это можно считать одной из высших форм интеллектуальной рефлексии.

Невычислимость также заставляет нас пересмотреть идею полной автоматизации познания. Если познание представить как последовательность вопросов, а каждый вопрос — как задачу, которую должен решать алгоритм, возникает надежда однажды создать универсальную машину познания. Но теория вычислимости говорит, что такой идеал требует гораздо более осторожной формулировки.

Можно автоматизировать огромные области рассуждения.

Можно создавать системы, которые сами строят доказательства.

Можно автоматически проверять корректность программ.

Можно создавать машины, способные генерировать новые гипотезы.

Можно строить интеллектуальные системы, которые превосходят человека в отдельных областях.

Но из этого не следует существование единой вычислительной процедуры, способной решать любую формально поставленную задачу.

Именно здесь проходит граница между универсальным интеллектом и универсальным решателем. Возможно, чрезвычайно мощный интеллект способен переключаться между различными методами, создавать новые представления и изменять собственные стратегии. Но даже такая гибкость не превращает невычислимую задачу в вычислимую.

Если существует доказанная невозможность универсального алгоритма, интеллект может изменить вопрос, ограничить область, использовать дополнительные знания или доказать невозможность. Но он не может просто «подумать достаточно сильно» и получить несуществующий алгоритм.

Это очень важный урок для будущего искусственного интеллекта. Вопрос не только в том, насколько большим станет количество параметров модели, сколько вычислений она сможет выполнять или насколько быстро будет обучаться. Гораздо важнее понимать, какие классы задач принципиально доступны алгоритмическим системам, а какие требуют изменения самой постановки.

Сверхинтеллект может быть сильнее человека почти во всех практических задачах и всё равно не быть универсальным решателем всех математических проблем.

Это не уменьшает его мощь. Наоборот, именно здесь становится возможным более зрелое определение сверхинтеллекта: не как существа, которому доступны абсолютно все ответы, а как системы, способной намного глубже человека исследовать пространство возможного, невозможного, доказуемого, недоказуемого, вычислимого и невычислимого.

Такой интеллект может знать не только ответы, но и карту границ между различными классами вопросов.

И эта карта сама по себе становится формой знания.

Можно даже предположить, что по мере развития интеллекта значение знания о невозможности будет возрастать. Для человека огромная часть интеллектуальной деятельности связана с поиском решения. Для более мощной системы значительная часть деятельности может быть связана с классификацией пространства задач: какие задачи решаются напрямую, какие требуют новых методов, какие зависят от дополнительных предположений, какие имеют только приближённые решения, а какие принципиально не имеют универсального алгоритма.

Это уже не просто вычислительная мощность. Это метаинтеллект — способность рассуждать о собственных способах рассуждения.

Но и здесь мы должны помнить урок предыдущих глав. Как только интеллект начинает анализировать собственные методы, возникает самоссылка. Если он пытается создать универсальную систему, которая полностью классифицирует все возможные интеллектуальные процедуры, она сама становится объектом анализа. А значит, вопрос о пределе вычислимости вновь возвращается на более высокий уровень.

Возможно, именно поэтому границы математического разума не образуют одну стену. Они образуют иерархию. За одной границей обнаруживается другая. За вычислимостью возникает вопрос о доказуемости. За доказуемостью — вопрос о выборе оснований. За основаниями — вопрос об интерпретации. За интерпретацией — вопрос о самом наблюдателе, который осуществляет интерпретацию.

Чем глубже мы исследуем интеллект, тем меньше он похож на машину, движущуюся по заранее построенной дороге. Он начинает напоминать систему, которая одновременно строит карту, исследует территорию и обнаруживает ограничения самой карты.

Именно поэтому невычислимость нельзя воспринимать как техническую деталь компьютерной науки. Она представляет собой одно из фундаментальных предупреждений против идеи абсолютной алгоритмизации разума.

Можно представить интеллект, который в миллиард раз быстрее человеческого. Можно представить интеллект, который одновременно анализирует миллиарды гипотез. Можно представить интеллект, способный самостоятельно создавать новые языки программирования, математические теории и архитектуры вычислений. Но если он остаётся алгоритмической системой, то существование невычислимых задач никуда не исчезает.

Увеличение интеллекта расширяет область достижимого.

Оно не обязано уничтожать саму границу достижимости.

И здесь мы приходим к ещё более тонкому вопросу. Если существуют невычислимые задачи, означает ли это, что за пределами вычислимого находится какое-то единое царство, которое можно назвать «невычислимым»?

Нет.

Невычислимость не образует простой альтернативный мир. Существуют разные степени и формы алгоритмической недоступности. Одни объекты могут быть вычислимы относительно дополнительных данных. Другие становятся доступными при расширении исходной системы. Некоторые задачи связаны с более высокими уровнями математической сложности. В теории вычислимости существует целая иерархия таких различий.

Это означает, что слово «невозможно» само требует анализа.

Невозможно каким алгоритмом?

Невозможно относительно каких исходных данных?

Невозможно для какого класса входных объектов?

Невозможно при каком требовании универсальности?

Невозможно доказать или невозможно вычислить?

Невозможно в принципе или невозможно при имеющихся ресурсах?

Интеллектуально зрелое мышление не смешивает эти утверждения.

Именно в этом заключается один из важнейших уроков настоящей главы. Граница разума начинается не там, где он сталкивается с трудной задачей. Она начинается там, где он перестаёт различать разные виды невозможности.

Можно бесконечно увеличивать вычислительные мощности и всё равно не получить универсального алгоритма, которого не существует.

Можно бесконечно увеличивать объём данных и всё равно не получить доказательства, которого нет в данной системе.

Можно бесконечно усложнять модель и всё равно не получить абсолютного знания о системе, если сама модель становится частью объекта анализа.

Именно поэтому следующий шаг должен быть ещё более радикальным.

До сих пор мы говорили о невозможности существования универсальных алгоритмов. Но если универсальный алгоритм невозможен, можно ли хотя бы для каждой конкретной программы определить, завершится ли она, используя достаточно умный анализ?

И здесь возникает знаменитая проблема, которая делает границу вычислимости особенно наглядной: **парадокс остановки**.

В следующей главе мы рассмотрим его уже не как короткий пример невычислимости, а как один из ключей к пониманию того, почему вычислительная система не может создать универсальный механизм, полностью предсказывающий поведение всех программ. Мы увидим, что проблема остановки — это не просто технический результат информатики. Она показывает нечто более глубокое: иногда система не может заранее вычислить собственное будущее, потому что попытка создать универсального предсказателя позволяет построить ситуацию, которая разрушает саму возможность такого предсказания.

Глава 8. Парадокс остановки

В предыдущей главе мы подошли к одной из наиболее глубоких границ алгоритмического мышления: существуют задачи, для которых невозможно создать универсальную вычислительную процедуру. Теперь необходимо рассмотреть эту границу более внимательно. Если невычислимость является общим свойством определённого класса задач, то возникает естественный вопрос: можем ли мы увидеть механизм этой невозможности настолько ясно, чтобы понять, почему она не устраняется даже принципиально более мощным компьютером?

Проблема остановки даёт на этот вопрос почти идеальный ответ. Она касается чрезвычайно простого вопроса: если нам известны программа и её входные данные, можем ли мы заранее определить, завершится ли программа или будет выполняться бесконечно?

На бытовом уровне этот вопрос кажется элементарным. Мы можем посмотреть на программу, запустить её, понаблюдать за поведением, найти цикл, обнаружить условие завершения. Для многих конкретных программ ответ действительно находится очень легко. Иногда достаточно нескольких секунд анализа. Иногда программа настолько очевидна, что ответ можно дать, не запуская её вообще.

Но проблема остановки ставится гораздо строже. Нас интересует не возможность решить вопрос для отдельных программ. Нас интересует существование универсального метода, который получает на вход описание любой программы и соответствующие входные данные и гарантированно определяет, завершится ли выполнение этой программы.

Именно слово «любой» превращает простой вопрос в фундаментальную проблему.

Если существует универсальный анализатор остановки, то он должен работать не только с простыми программами, поведение которых очевидно. Он должен одинаково надёжно обрабатывать любые допустимые программы: короткие и длинные, простые и сложные, созданные человеком и автоматически сгенерированные, математические, игровые, самореферентные, оптимизирующие собственный код и анализирующие другие программы.

Ему нельзя разрешить ошибаться.

Ему нельзя отвечать «не знаю».

Ему нельзя зависать вместе с анализируемой программой.

Ему необходимо для каждого случая за конечное время выдать один из двух ответов: программа остановится или программа не остановится.

И вот здесь возникает фундаментальная невозможность.

Важно сразу отметить: название «парадокс остановки» несколько обманчиво. Перед нами не парадокс в обычном смысле слова, подобный логическому противоречию. Это задача, показывающая невозможность существования определённого универсального алгоритма. Её сила состоит не в том, что мы сталкиваемся с загадочным противоречием, а в том, что можем строго показать: предполагаемый универсальный предсказатель неизбежно будет ошибаться хотя бы на одной специально сконструированной программе.

Чтобы понять идею, представим гипотетическую машину анализа. Назовём её Предсказателем. Она получает описание любой программы и её входа. После анализа она должна сообщить, что произойдёт.

Если программа завершится, Предсказатель сообщает: «остановится».

Если программа будет работать бесконечно, Предсказатель сообщает: «не остановится».

Предположим пока, что такая машина существует. Не будем сразу спорить с этим предположением. Напротив, дадим ей максимальные возможности. Пусть она чрезвычайно быстра, обладает огромной памятью и использует любые доступные алгоритмические методы. Пусть её создатели потратили на неё столетия исследований. Пусть она намного умнее любого современного анализатора программ.

Теперь построим другую программу.

Эта программа получает описание некоторой программы и передаёт его Предсказателю. Затем она делает ровно противоположное тому, что предсказал Предсказатель.

Если Предсказатель говорит, что исходная программа остановится, новая программа входит в бесконечный процесс и не останавливается.

Если Предсказатель говорит, что исходная программа не остановится, новая программа немедленно завершает работу.

Пока это кажется просто хитрой конструкцией. Но теперь самое интересное.

Представим, что мы передаём самой этой новой программе её собственное описание.

То есть программа получает информацию о самой себе и спрашивает Предсказателя, что произойдёт с ней самой.

Если Предсказатель отвечает: «эта программа остановится», программа обязана сделать обратное и не остановиться.

Если Предсказатель отвечает: «эта программа не остановится», программа делает обратное и останавливается.

Предсказатель в любом случае ошибается.

Значит, универсального Предсказателя с требуемыми свойствами существовать не может.

Это рассуждение кажется почти детской логической игрой только до тех пор, пока не осознаешь его последствия. Здесь не удалось найти слабое место конкретного компьютера. Не было доказано, что современный процессор недостаточно быстр. Не было показано, что нужно больше памяти. Не было обнаружено, что программисты ещё не придумали правильный язык программирования.

Было показано невозможное существование универсального алгоритма.

Это совершенно другой тип результата.

Представим, что мы строим компьютер, который в тысячу раз быстрее другого. Для огромного числа задач это принципиально меняет практическую ситуацию. Программа, работавшая десять часов, будет работать несколько секунд. Затем мы строим машину ещё в миллион раз мощнее. Некоторые задачи, казавшиеся недоступными, становятся практически решаемыми.

Но с проблемой остановки увеличение мощности не меняет фундаментального положения. Если универсальный анализатор невозможен, то невозможно получить его простым увеличением скорости.

Можно построить более быстрый компьютер.

Можно построить компьютер с большей памятью.

Можно создать более эффективный алгоритм анализа.

Можно разработать специализированный метод для определённого класса программ.

Можно использовать несколько компьютеров одновременно.

Можно применять искусственный интеллект для поиска закономерностей в коде.

Можно создавать системы формальной верификации.

Но нельзя получить универсальную процедуру, которая безошибочно определяет остановку любой возможной программы.

Именно здесь особенно ясно проявляется различие между ограничением ресурсов и логическим ограничением.

Ограничение ресурсов говорит: «Мы пока не можем решить эту задачу».

Ограничение алгоритмического типа говорит: «Не существует универсального метода, который решает эту задачу для всех случаев».

Первое ограничение может исчезнуть благодаря технологическому прогрессу.

Второе не исчезает от увеличения мощности в рамках того же вычислительного принципа.

Это различие имеет огромное значение для понимания будущего искусственного интеллекта. Когда говорят о будущих машинах, часто предполагают почти линейную картину развития: больше вычислений, больше памяти, больше данных, более совершенные архитектуры — и постепенно исчезают все ограничения, которые сегодня кажутся фундаментальными.

Но теория вычислений показывает, что такое представление неверно.

Увеличение мощности действительно способно перемещать множество практических границ. Однако оно не отменяет логических границ самого вычисления.

Чтобы понять это ещё глубже, необходимо рассмотреть, почему Предсказатель оказывается бессилен. Причина не в том, что он недостаточно хорошо изучил программу. Напротив, предположим, что он знает её абсолютно точно. Он знает каждую инструкцию, каждый переход, каждое условие. Он может полностью анализировать её структуру.

Проблема возникает потому, что анализируемая программа получает возможность использовать информацию о собственном предсказании.

Это принципиально.

Если предсказатель находится полностью вне объекта и его прогноз никак не влияет на объект, ситуация может быть иной. Но если объект способен получить описание прогноза и построить своё поведение в зависимости от этого прогноза, возникает обратная связь.

Предсказание становится частью причины того поведения, которое оно пытается предсказать.

Именно здесь проблема остановки связывается с темой самоссылки, которую мы рассматривали в предыдущих главах.

В обычном представлении прогноз — это информация о будущем. Мы предполагаем, что существует объект, который действует, а затем внешний наблюдатель сообщает нам, что произойдёт.

Но что если объект узнаёт прогноз заранее?

Если объект способен изменить своё поведение в зависимости от прогноза, прогноз перестаёт быть нейтральным описанием. Он становится элементом самой системы.

Для простых физических объектов это может не иметь значения. Если астроном сообщает, что через час будет солнечное затмение, само знание этого прогноза не изменяет движение планет. Но для интеллектуального агента ситуация может быть совершенно другой.

Представим, что человеку сообщают: «Мы абсолютно уверены, что завтра вы выберете вариант А».

Человек может сказать: «Тогда я специально выберу вариант Б».

Сам прогноз изменил условия, относительно которых прогноз строился.

Конечно, человеческое поведение значительно сложнее математической программы, и этот пример нельзя механически отождествлять с проблемой остановки. Но структурная аналогия важна: система, способная реагировать на собственное предсказание, может разрушить возможность некоторых форм универсального предсказания.

Теперь можно сделать следующий шаг.

Проблема остановки показывает не просто невозможность предсказывать программы. Она показывает предел способности системы строить универсальную модель поведения других вычислительных систем, если эти системы достаточно выразительны и способны участвовать в самоссылочной конструкции.

Это означает, что «иметь полное описание правил» и «уметь заранее вычислить все последствия правил» — разные вещи.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, купив полную легальную версию на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.