ДУМАЙТЕ ЛУЧШЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УПРАВЛЯЙТЕ МЫШЛЕНИЕМ,
А НЕ ЗНАНИЯМИ

РЭН АСАХИНА

Рэн Асахина

Думайте лучше искусственного интеллекта. Управляйте мышлением, а не знаниями

https://litres.ru/74448757
SelfPub; 2026

Аннотация

ИИ уже умеет знать больше вас. Значит ли это, что он должен думать за вас? Эта книга ставит на метапознание способность замечать, как именно вы мыслите, где ошибаетесь, чего не знаете и почему доверяете собственному выводу. Вместо гонки за фактами автор учит калибровать уверенность, видеть неизвестные неизвестности, задавать сильные вопросы, управлять вниманием, проверять предположения и учиться быстрее. Внутри — аудиты уверенности и внимания, pre-mortem, «пять почему», работа с Сократическими вопросами и примеры того, как ИИ создаёт иллюзию компетентности. Это книга не о том, как победить машину объёмом знаний. Она о том, как сохранить то преимущество, которое действительно остаётся человеческим: способность управлять собственным мышлением.

Рэн Асахина
Думайте лучше искусственного интеллекта. Управляйте мышлением, а не знаниями

Научная фантастика уже давно рассказала нам

Тед Чанг в своем рассказе «Выдох» (Exhalation, 2008) описал анатома. Он жил в механическом мире, движимом атмосферным давлением, где все обитатели функционировали за счет аргона в лёгких. Однажды он заметил, что часы замедлились — не одни часы, а все. Никто больше не обратил на это внимания. Он начал подозревать: не часы замедлились, а мы ускорились — или, точнее, изменился наш мозг.

Чтобы проверить эту гипотезу, он сделал то, чего в его мире еще никто не делал: вскрыл себе череп и, используя зеркало, наблюдал за тем, как он думает, одновременно размышляя.

Оно увидело, как золотые лепестки перелистываются в потоке газа — это и был физический процесс мышления. Оно осознало, что аргон, питающий мысли, иссякает, и вся цивилизация обречена на тепловую смерть. Но оно не впало

в панику. Оно использовало оставшееся время, чтобы записать всё наблюдаемое в виде записки, оставленной для тех, кто в будущем, возможно, откроет для себя существование этого мира.

Этот роман не о конце света. В нём речь идёт о том, что, когда ты способен наблюдать за собственным мыслительным процессом, ты обретаешь фундаментальную свободу — даже перед лицом предопределённого финала ты по-прежнему обладаешь способностью понимать и выбирать.

«Думать и одновременно наблюдать за тем, как ты думаешь» — именно так определяется метапознание.

Писатели-фантасты снова и снова обращаются к этой теме, поскольку это самый сущностный вопрос, возникающий у человека при столкновении с интеллектом, превосходящим его собственный.

В романе Дэниела Кейса «Цветы для Элджернона» (*Flowers for* Algernon, 1966) умственно отсталый Чарли благодаря операции обретает сверхчеловеческий интеллект. Самым мучительным открытием для него после обретения ума становится не сложность мира, а то, что он наконец-то может увидеть то, чего раньше не замечал — в том числе насмешки окружающих, ограниченность собственного познания и то, что все это в конечном итоге будет утрачено. Когда его интеллект достиг пика перед началом упадка, он написал: «Одно из самых важных вещей, которые я теперь пони-

маю, — это то, что я всё это время не знал, чего я не знаю». Это не история об интеллекте. Это история об осознании границ познания — теме третьей главы этой книги.

В романе Лю Цисиня «Три тела» (2008) основная логика проекта «Стеноходцы» заключается в следующем: жители планеты «Три тела» могут считывать всю информацию о человечестве, но не могут проникнуть в сам процесс человеческого мышления. Оружием «Стеноходцев» являются не больше знаний или более мощные вычислительные ресурсы, а стратегическое использование собственного мышления — о чем думать, о чем не думать, когда и в каком уголке какого уровня мышления скрывать истинные намерения. Ло Цзи в конечном итоге одержал победу не потому, что «знал» больше, чем жители «Трех тел», а потому, что лучше всех понимал, как он сам мыслит в этой игре. Проект «Стеноходцы» — это экстремальный эксперимент в области метапознания.

В рассказе Азимова «Последний вопрос» (*The Last* Question, 1956) на протяжении нескольких триллионов лет люди постоянно задают всё более мощным суперкомпьютерам один и тот же вопрос: «Можно ли обратить вспять энтропию?» Каждое поколение машин отвечает: «Недостаточно данных, не могу ответить». Только после наступления тепловой смерти Вселенной машина наконец нашла ответ — но людей, которые могли бы его услышать, уже не осталось. Настоящая тема этого рассказа — не термодинамика. В нём задаётся вопрос: когда машины смогут ответить на всё, кто бу-

дет решать, какие вопросы стоит задавать?

О чём эта книга

Общая нить, проходящая через все эти рассказы: самое глубокое преимущество человечества никогда не заключалось в том, чтобы «знать больше», а в осознании и контроле над собственным познанием.

Тед Чанг в «Анатомии» ученый, вскрывающий себе череп, чтобы посмотреть на себя изнутри. Чарли, достигший пика интеллектуального развития, осознал, чего он раньше не знал. Ложи, превративший само мышление в оружие. Люди Азимова, которые перед лицом машин так и не задали тот единственный, по-настоящему важный вопрос.

ИИ становится суперкомпьютером нашей эпохи. С каждым годом он берет на себя все больше задач, связанных с «мышлением»: анализ, поиск, генерация, прогнозирование, оптимизация. Большинство людей реагируют на это тем, что «учатся пользоваться ИИ» — используют более эффективные подсказки и оптимизируют рабочие процессы. Это, конечно, важно. Но при этом упускается из виду более фундаментальный вопрос:

какие способности «мыслить» вам по-прежнему нужны, когда ИИ думает быстрее вас и знает больше вас?

Ответ — метапознание, то есть размышление о самом процессе мышления. Речь идет не о том, о чём думать, а о том, чтобы наблюдать за тем, как вы думаете, решать, о чём вам следует думать, и нажимать на кнопку «пауза», когда

нужно усомниться.

Эта книга не учит вас, как использовать ИИ. Эта книга учит вас, как сохранять ясность ума в эпоху ИИ.

Что вы извлечёте из книги

Аспекты

Конкретные выводы

Навыки

Осознание границ познания (понимание того, чего вы не знаете), постановка точных вопросов, регулирование внимания, отбор аналогий из разных областей, оценка ценности

Результаты

Список упражнений по метапознанию ( ), шаблон аудита внимания, структура семинара по междисциплинарным связям, эскиз личной метапознавательной операционной системы

Психологический настрой

Перестать стремиться знать больше, чем ИИ, а стремиться задавать вопросы лучше, чем ИИ; уметь использовать неизвестное; активно регулировать когнитивную нагрузку, не позволяя информации вести себя за собой

Обзор глав

Обзор глав

Глава
Название (китайский)
Название (англ.)
Основные вопросы

01
Инверсия
The Inversion
Когда ИИ знает больше вас, что останется?

02
Познание и метапознание
Cognition and Metacognition
Почему «думать» и «наблюдать за тем, как ты думаешь» — это две разные вещи?

03
Осознание того, чего ты не знаешь
Осознание своих пробелов в знаниях
Самая недостающая способность ИИ: осознание границ познания

04
Искусство задавать вопросы
Искусство задавать вопросы

Ваша метакогнитивная ОС
Рамки от концепции до повседневной практики

10
Будущее человеческого познания
Будущее человеческого познания
Куда ведёт нас метапознание

Как читать

Рекомендуемый маршрут

Маршрут A (практический подход): 01 → 04 → 05 → 09. Перейдите сразу к главам «Переворот», «Вопросы», «Внимание» и «Операционная система» — эти четыре главы позволят создать применимую на практике метакогнитивную модель.

Маршрут B (приоритет понимания): 01 → 02 → 03 → 07 → 10. Сначала разберитесь в «Что осталось», «Двухуровневая структура», «Границы неизвестного», «Способность к суждению» и «Направления развития», а уже потом решите, стоит ли углубляться в методы.

Маршрут C (полное прочтение): по порядку глав 01–10. Подходит читателям, желающим систематически построить систему метапознания и имеющим время для выполнения упражнений.

Можно пропустить

Если вас не интересует концептуальная основа когнитивной науки, раздел 02 можно прочитать бегло — однако ре-

комендуется ознакомиться хотя бы с двухслойной моделью из подраздела 2.2, поскольку в последующих разделах к ней будут неоднократно обращаться.

Если вас не интересует вопрос «почему ИИ не может сделать то-то и то-то», можете пробежаться глазами по первой половине раздела 07.

Если у вас уже сформировались навыки управления вниманием, раздел 05 можно прочитать бегло.

Раздел 10 носит прогнозный характер; если вас интересуют только методы, его можно пропустить.

Метод трёхкратного чтения

Первое чтение: быстро просмотрите заголовки и подзаголовки, чтобы составить общее представление о книге (около 30 минут).

Второе чтение: внимательно прочитайте разделы, связанные с вашими проблемами, делайте пометки и записывайте вопросы.

Третье прочтение: с учетом структуры из раздела 09 объедините отмеченный материал в личный метакогнитивный список.

Условия использования книги

Условия
Пояснения

Диаграммы

Использование векторных диаграмм SVG (созданных в draw.io) для отображения структур, процессов и взаимосвязей; к каждой главе прилагаются диаграммы на китайском и английском языках, расположенные в каталоге images/.

Язык

В китайском издании преобладает упрощённый китайский; при первом упоминании собственных названий приводится их английский вариант.

Форматирование

Ключевые понятия при первом упоминании выделяются жирным шрифтом; инструкции представлены в виде нумерованных списков; примеры и цитаты оформлены в виде блочных цитат.

Двуязычность

Названия глав приводятся на китайском и английском языках; при первом упоминании ключевых терминов дается их английский вариант.

«Анатом» Теда Чжана в конце концов превратился в письмо к неизвестному читателю. Эта книга — тоже письмо, адресованное вам, стоящему на пороге эры искусственного интеллекта. Если вам интересно, к чему в конечном итоге приведет развитие человеческого познания, перейдите

к десятой главе. Если же вы больше хотите узнать, что делать прямо сейчас, начните с первой главы.

Глава 1: Инверсия
The Inversion

Основной вопрос: что останется, когда ИИ будет знать больше, чем вы? / When AI knows more than you, what remains?

1.1 Вступительная история: шесть несуществующих судебных прецедентов

В июне 2023 года судья Федерального суда Южного округа Нью-Йорка П. Кевин Кастел в одном из публичных постановлений написал редкую фразу: «Наш суд никогда не сталкивался с такой беспрецедентной ситуацией — адвокат с тридцатилетним стажем представил суду шесть совершенно несуществующих судебных прецедентов».

Суть дела несложна. Адвокат Стивен Шварц представлял интересы клиента в иске против авиакомпании «Авианка» (Avianca) по делу *Mata v. Avianca Inc.* Для подготовки юридического заключения он использовал ChatGPT для поиска судебных прецедентов. ИИ предоставил ему шесть ссылок — с названиями дел, судами, в которых они рассматривались, датами вынесения решений и кратким изложением сути решений — со всеми необходимыми деталями. Шварц не проверил в Westlaw или какой-либо другой юридической базе данных, действительно ли эти прецеденты существуют. Он сразу же включил их в официальные документы, поданные в федеральный суд.

Адвокат противоположной стороны, не найдя этих дел в базах данных, подал возражение судье. Судья потребовал от Шварца предоставить тексты этих судебных прецедентов. Шварц обратился к ChatGPT с вопросом: «Ты уверен, что эти прецеденты подлинные?» ИИ ответил: «Да, они подлинные». Он также представил этот ответ суду.

Все шесть судебных прецедентов были выдуманы ИИ. Названия дел были вымышленными, суды — вымышленными, даты — вымышленными. Ни один из них не фигурировал в базе данных федеральной судебной практики. Шварц и его коллега Питер ЛоДука были публично наказаны судьей, оштрафованы на 5000 долларов и обязаны направить письма с извинениями каждому судье, «упомянутому» в вымышленных прецедентах.

Суть этого дела не в том, что ИИ допустил ошибку. Галлюцинации крупных языковых моделей — генерация контента, выглядящего правдоподобно, но являющегося полностью вымышленным — уже давно являются известной проблемой в технологическом сообществе. Настоящий повод для размышлений — человеческий фактор: адвокат с тридцатилетним стажем, получив результат работы ИИ, ни разу не подумал: «Мне нужно это проверить».

Дело не в том, что он не разбирался в праве. Ему не хватало более фундаментальной способности: осознания собственного познавательного процесса. Он не задал себе вопроса: «Насколько моё понимание этих судебных прецеден-

тов основано на моём профессиональном суждении, а насколько — на слепом доверии к машине?» Он не сумел найти ту точку между доверием и скептицизмом, где следовало остановиться. Он принял уверенность ИИ за свою собственную.

Дело Шварца — не единичный случай, а сигнал, который свидетельствует о том, что в отношениях между человеком и машиной в сфере знаний происходят структурные изменения. На протяжении сотен лет наша профессиональная идентичность основывалась на одном неявном допущении: человек знает больше, чем машина. Врач разбирается в диагностике лучше, чем поисковая система, адвокат — в защите лучше, чем база данных, а аналитик — в тенденциях лучше, чем электронная таблица. Но когда ИИ способен за считанные секунды сгенерировать, казалось бы, полное юридическое заключение, четко структурированный отраслевой отчет или логически связный фрагмент кода, это допущение перестает быть надежным.

Асимметрия знаний меняется на противоположную. Раньше было так: «Я знаю то, чего ты не знаешь», а теперь стало: «Машина, возможно, знает больше, чем я — но я не уверен, верно ли то, что она знает». В этой новой ситуации «знать больше» больше не является ключевым преимуществом человека. Настоящей редкостью является способность, которую ИИ до сих пор с трудом может воспроизвести: осознавать, что ты знаешь, а чего не знаешь, понимать,

когда следует доверять, а когда — сомневаться. Эта способность называется метапознанием (metacognition). Речь идет не о грандиозном вопросе «заменит ли ИИ человека», а о личном поиске ответа на вопрос: «В эпоху, когда ИИ способен производить больше, чем ты, какую незаменимую ценность ты все еще можешь предложить?».

Реальный отклик: дело Шварца — не единственный знаковый случай «переворота в соотношении знаний человека и машины». В том же году работа *«Théâtre D'opéra Spatial»,* созданная Джейсоном Алленом с помощью инструмента искусственного интеллекта Midjourney, получила первую премию в номинации «Цифровое искусство» на художественной выставке в Колорадо — при этом члены жюри не знали, что произведение было сгенерировано ИИ. Это событие вызвало глобальную дискуссию о границах творчества: когда машина способна создавать произведения, превосходящие работы большинства человеческих участников, что означает статус «творца»? От судов до галерей, от судебных прецедентов до картин — переворачивание асимметрии знаний — это не теоретическое предсказание, а реальность, которая одновременно происходит во всех сферах.

1.2 Основная теза: переворачивание асимметрии знаний

Отношения между человеком и машиной долгое время строились на одном неявном допущении: человек знает больше, чем машина. Машина выполняет, человек принима-

ет решения; машина вычисляет, человек судит; машина хранит, человек творит. Это допущение лежало в основе промышленной и информационной революций, а также сформировало наше представление об «экспертах». Врач разбирается в диагностике лучше, чем база данных, адвокат — в защите лучше, чем поисковая система, а аналитик — в тенденциях лучше, чем Excel.

В последние десять лет это предположение постепенно опровергается.

Шахматы, го, написание текстов, программирование, рыночный анализ, юридический поиск — в одной области за другой машины начинают понимать больше, чем люди. Под «пониманием» здесь подразумевается не только объем хранимой информации, но и синтез, логическое мышление, генерация. Один подключенный к сети ИИ способен за несколько секунд обобщить литературу, накопленную человечеством за несколько столетий; опытный эксперт, даже посвятив этому всю свою жизнь, сможет внимательно изучить лишь небольшую её часть. Когда ИИ сможет свободно составлять отраслевые отчёты, генерировать код и отвечать на специализированные вопросы, тот барьер, на котором держались работники умственного труда — «я знаю то, чего ты не знаешь», — начнёт рушиться слой за слоем.

Бринйольфссон и МакАфи в книге *«Вторая эра машин»* (, 2014) отмечают: ускорение технического прогресса переписывает соотношение способностей человека и маши-

ны. Это не постепенная корректировка, а структурная инверсия (structural inversion). Когнитивные высоты, когда-то принадлежавшие исключительно человеку, шаг за шагом захватываются машинами. В книге приводится образное сравнение: мы «соревнуемся с машинами», но правила игры изменились. Соревнование больше не заключается в том, кто знает больше, а в том, кто знает, что он знает, а чего не знает.

Таким образом, на поверхность выходит ключевой вопрос: когда асимметрия знаний переворачивается, в чём же заключается ценность человека?

Дело не в том, чтобы «знать больше» — в любой специализированной области вам будет трудно превзойти ИИ. Дело не в том, чтобы «делать быстрее» — по скорости и масштабам ИИ подавляет человека. Ответ кроется в способности, которую ИИ до сих пор не может воспроизвести: метапознании (metacognition), то есть осознании, контроле и регулировании собственных когнитивных процессов.

Метапознание интересует не то, что вы знаете, а то, как вы думаете. Знаете ли вы, что вы знаете, а чего не знаете? Знаете ли вы, когда следует доверять собственному суждению, а когда — подвергать его сомнению? Именно эта способность является слабым местом ИИ. В этой главе будут рассмотрены три вопроса: почему произошло это переворачивание, почему метапознание является когнитивной операционной системой человека в эпоху ИИ (обновляемой инфраструктурой, а не статической линией обороны), и как вам

начать осознавать и применять его.

1.3 Доказательства и мыслительные эксперименты

1.3.1 Краткая история переворота асимметрии знаний

Перелом произошёл не за одну ночь. Если оглянуться на соперничество человека и машины в вопросе «кто знает больше», можно увидеть чёткую эволюционную кривую. За каждым важным этапом стоит очередное отступление человечества от границ того, «чего машина не может сделать».

1997 год, шахматы. «Депл Блу» победил Каспарова. Позже Каспаров сказал, что почувствовал нечто вроде «инопланетного интеллекта». Но тогда все сходились во мнении: правила шахмат фиксированы, пространство поиска ограничено, а машина просто применяет метод перебора. Пространство состояний в игре в го гораздо больше, чем в шахматах, а для писательского и творческого труда нужны интуиция и эмоции. Человечеству пока не о чем беспокоиться.

2016 год, го. AlphaGo победила Ли Се-ши. Ещё более поразительными оказались некоторые её ходы: профессиональные игроки сначала сочли их неудачными, но после анализа партии обнаружили в них глубокую стратегию. Машина не только умела рассчитывать, но и демонстрировала способность делать выбор, похожий на «интуицию». Общественное мнение начало колебаться. Однако такие области, как писательское творчество и принятие бизнес-решений, связанные с открытым контекстом, по-прежнему считались прерогати-

вой человека. Граница «машина понимает правила, человек понимает мир» по-прежнему существует.

2020–2023 годы: письмо и программирование. Появление GPT-3 и последующих моделей. Повсеместное распространение помощников по программированию, инструментов для создания резюме и переводчиков. Программисты начали использовать Copilot для автодополнения больших фрагментов кода, а авторы — генерировать черновики с помощью ИИ, а затем дорабатывать их. Впервые работники умственного труда массово убедились, что «то, что написано ИИ, ничем не хуже того, что написал бы я». Граница снова сдвинулась: даже в такой открытой области, как естественный язык, машины могут создавать конкурентоспособный контент.

С 2023 года по настоящее время: анализ и отчетность. ИИ способен быстро интегрировать данные по заданной теме и генерировать структурированные отчеты. Аналитики, исследователи и консультанты обнаруживают: на уровне создания контента ИИ уже не уступает, а порой даже превосходит человека. Разница зачастую проявляется лишь в тонком понимании потребностей, оценке границ и метакогнитивной калибровке качества. Предположение о том, что «человек знает больше, чем машина», уже не подтверждается во всё большем числе вертикальных областей. Переворот из исключения превратился в норму.

Рис. 1-1: Переворот в асимметрии знаний

Этот рисунок обобщает логику такого переворота: в до-ИИ-эпоху преимущество было за человеческими знаниями, в эпоху ИИ — за знаниями машин, а метакогнитивный поворот означает, что преимущество человека смещается с «что он знает» на «как он осмысливает то, что знает».

Этапы
Относительное преимущество человека
Относительное преимущество машин

Доэра ИИ
Хранение знаний, профессиональные суждения, творчество
Скорость вычислений, объем памяти

Эпоха ИИ (настоящее время)
Метапознание, способность к принятию решений, ответственность
Интеграция знаний, скорость генерации, сопоставление шаблонов

После метакогнитивного перелома
Как думать, как принимать решения, как корректировать
То же самое, причем в постоянно расширяющемся масштабе

1.3.2 Мысленный эксперимент: если завтра ИИ

усвоит все ваши профессиональные знания

Предположим, что со завтрашнего дня весь профессиональный багаж в вашей области — концепции, формулы, кейсы, передовой опыт — будет полностью усвоен ИИ. Он сможет, как и вы, отвечать на вопросы клиентов, составлять отраслевые отчеты, давать профессиональные рекомендации. У вас больше не будет никакого преимущества в плане «запаса знаний». Клиенты смогут обращаться напрямую к ИИ, и ответы будут не хуже ваших; начальник сможет поручить ИИ составлять отчеты, и, возможно, он будет делать это быстрее и лучше вас.

Какую ценность вы сможете предложить?

Большинство людей застрянут именно на этом. Наша профессиональная идентичность долгое время строилась на принципе «я знаю, а другие — нет». Врачи полагаются на диагностику, юристы — на знание законов, аналитики — на отраслевые знания. Как только эта основа исчезает, чувство идентичности рушится — когда «я знаю» больше не действует, возникает вопрос «кто я?».

Но если задуматься глубже, можно обнаружить некоторые ниши, которые ИИ заменить не сможет:

Вы знаете, чего вы не знаете. Вы умеете определять границы проблемы, понимаете, что выходит за рамки ваших текущих знаний и требует обращения за помощью или дополнительного исследования. ИИ склонен к излишней самоуверенности и с трудом признает: «В этом я не уверен». Он мо-

жет сгенерировать ответ, который выглядит полным, даже если он основан на ошибочных предпосылках или устаревших данных. А вы можете сказать: «В этом я не уверен, нужно проверить».

Вы знаете, когда стоит доверять ИИ, а когда — подвергать его сомнению. Есть ли в логической цепочке этого отчета разрывы? Надежны ли источники данных? Не зашли ли выводы слишком далеко? Текст, сгенерированный ИИ, часто звучит плавно и связно, но плавность не означает правильность. Люди, способные самостоятельно «нажать на тормоза» и поставить под сомнение качество результатов, в эпоху ИИ, напротив, становятся еще более редкими.

Вы знаете приоритеты задач. Среди бесчисленного множества дел, которые можно выполнить, какие действительно стоят того, чтобы ими заняться? А какие только кажутся срочными? ИИ может выполнять задачи, но ему трудно самостоятельно судить о том, «что стоит рассматривать как проблему». Клиент выдвигает десять требований, и вам нужно определить, какие три из них имеют стратегическое значение, а какие семь можно отложить. Такое суждение основано на понимании контекста бизнеса, а не просто на поиске информации.

Вы умеете принимать решения в условиях неопределённости. Информация неполна, модели ненадежны, но вам всё равно приходится принимать окончательное решение. Такое решение основано на суждении, а суждение, в свою очередь,

строится на «осознании границ собственного познания». Вы знаете, на что делаете ставку, знаете, насколько широк доверительный интервал, знаете, когда следует действовать осторожно, а когда можно пойти на риск.

Все эти способности имеют одну общую черту: все они связаны с метапознанием. Важно не то, сколько вы знаете, а то, как вы применяете свои знания, осознаете пределы своих знаний и понимаете, когда следует корректировать свою когнитивную стратегию.

1.3.3 Данные исследований: ИИ делает вас умнее, но не мудрее

В 2025 году Шариати и его коллеги опубликовали в журнале «*Computers in Human* Behavior» статью с очень удачным названием: «*AI Makes You Smarter but None the Wiser*». «None the wiser» — это английская идиома, означающая, что после совершения какого-либо действия человек по-прежнему находится в неведении. Исследователи использовали его для описания парадокса: ИИ позволяет вам лучше справляться с задачами, но не дает вам более ясного представления о себе.

Дизайн исследования был довольно простым. Участники выполняли задания по написанию текстов, анализу и принятию решений как с помощью ИИ, так и без него, а после завершения оценивали свои результаты (прогнозировали баллы, определяли своё место в рейтинге группы). Независимые эксперты выставляли оценки. Сравнивая «самооценку» с «фактическими баллами», можно было измерить метако-

гнитивную калибровку (metacognitive calibration), то есть насколько точно вы оцениваете свои результаты.

Результаты свидетельствуют о трёх моментах:

Результаты действительно улучшились. Участники, использовавшие ИИ, продемонстрировали значительно более высокое качество результатов, чем те, кто обходился без него. Это соответствует интуиции: при грамотном использовании ИИ отчеты действительно получаются лучше.

Однако самооценка, напротив, стала менее точной. У участников, использовавших ИИ, отклонение в оценке собственного уровня выполнения задания было больше, а кривая калибровки отклонялась от идеального состояния. После того как вы полагаетесь на результаты ИИ, вам становится сложнее точно ответить на вопрос «Как я, собственно, справился?».

Чем лучше вы умеете пользоваться ИИ, тем менее точна ваша самооценка. У людей с более высокой ИИ-грамотностью метакогнитивная неточность выражена сильнее. Это противоречит интуиции: можно было бы подумать, что чем лучше человек владеет инструментом, тем четче он понимает пределы его возможностей. Эмпирические данные показывают прямо противоположное.

Именно об этом говорится в названии «None the Wiser»: вы стали умнее, но не мудрее. Под мудростью здесь подразумевается трезвое осознание собственного когнитивного состояния. Использование ИИ, как раз, ослабляет это осозна-

ние.

Что это означает? Если мы будем полагаться на ИИ без рефлексии, то, повышая краткосрочную эффективность, мы рискуем подорвать свои ключевые долгосрочные преимущества. В некоторой степени само это чувство кризиса является метакогнитивным сигналом — по крайней мере, вы все еще задаетесь вопросом: «Что я на самом деле делаю?». Те, кто полностью полагается на ИИ, но больше не задается этим вопросом, возможно, идут по пути, ведущему к тому, что они станут «умнее, но менее осознанными».

1.3.4 Парадокс: чем выше уровень грамотности в области ИИ, тем менее точна самооценка

Исследование Шариати и других ученых выявило парадокс, противоречащий интуиции: чем чаще используется ИИ, тем сложнее точно оценить себя.

На первый взгляд это нелогично. Человек, умело использующий инструмент, должен лучше понимать, на что способен этот инструмент, а на что нет, и более точно различать «мой вклад» и «вклад инструмента». Однако эмпирические результаты показывают прямо противоположное. Почему?

Три возможных объяснения:

Неясность атрибуции. ИИ проникает во все этапы написания текстов, анализа и программирования, и человеку трудно отличить «это я придумал» от «это предложил ИИ». Результат становится смешанным продуктом, и ориентиры для самооценки стираются.

Иллюзия плавности. Тексты, сгенерированные ИИ, часто отличаются плавностью и связностью, читаются вполне убедительно. Такая плавность создает ощущение контроля: кажется, что вы сами поняли содержание. На самом деле вы, возможно, просто редактируете, а не создаете, но при этом чувствуете себя хорошо.

Иллюзия переноса способностей. Люди, часто использующие ИИ, могут ошибочно интерпретировать «я умею настроить ИИ так, чтобы он давал хорошие результаты» как «я сам обладаю высокими способностями». Первое — это способность к сотрудничеству, второе — способность действовать самостоятельно. Это разные вещи, но в повседневной практике их легко спутать.

Вывод из этого парадокса очевиден: суть метакогнитивных вызовов эпохи ИИ заключается в проблеме самопознания, а не в технических вопросах. Вам нужны не дополнительные навыки работы с ИИ, а более трезвое осознание: чем я занимаюсь? Какой вклад я вношу? Что я знаю, а чего не знаю?

Если это наблюдение останется верным, «научиться использовать ИИ» и «сохранять метакогнитивную ясность» станут противоположными тенденциями. Чем больше вы используете ИИ, тем лучше ваши результаты, но тем хуже ваша самооценка; не используя ИИ, ваша самооценка будет точнее, но результаты могут отставать. Выход заключается не в выборе одного из двух, а в активном формировании привыч-

ки к метапознанию. Используя результаты работы ИИ, сознательно задавайте себе вопросы: «В чём заключается мой вклад? Проверил ли я качество этого результата? В какой степени я полагаюсь на ИИ? Сам по себе этот вопрос является сопротивлением метакогнитивной неточности.

1.3.5 Что такое метапознание: осознание того, что вы делаете со «своими знаниями»

Концепция метапознания (metacognition) была систематически изложена когнитивным психологом Джоном Флавеллом (John Flavell) в 1979 году. В своей статье «Метапознание и когнитивный мониторинг» (Metacognition and Cognitive Monitoring), опубликованной в журнале «Американский психолог», он определил метапознание как «познание о познании» (cognition about cognition): осознание, контроль и регулирование собственных мыслительных процессов. Флавелл исследовал когнитивное развитие детей, но его концептуальная модель носит универсальный характер. Метапознание — это не привилегия избранных гениев, а часть когнитивной системы человека. Просто мы привыкли к тому, как оно работает, и только какой-то импульс (например, ИИ) заставляет нас переосмыслить это.

Флавелл разбивает метапознание на четыре взаимосвязанных компонента:

Метакогнитивное знание. Знания о когнитивных задачах, субъекте познания и когнитивных стратегиях. Например: вы знаете, что заучивание списка легче забывается, чем пони-

мание концепции; вы знаете, что в математике у вас больше уверенности, чем в языке; вы знаете, что распределенная практика эффективнее, чем интенсивная подготовка в последнюю минуту. Эти знания помогают вам прогнозировать сложность задачи, выбирать стратегию и формировать ожидания.

Метакогнитивный опыт. Эмоции и ощущения, возникающие в процессе познания. Чувство «застрял» при чтении сложного отрывка, внезапное «озарение» при написании доклада, ощущение «что-то не так» при принятии решения. Эти переживания часто остаются вне поля сознательного внимания, но они являются важными сигналами метакогнитивного контроля, подсказывающими вам: «здесь нужно что-то скорректировать».

Цель. Какова цель вашей текущей когнитивной деятельности? Без чёткой цели метапознание не может эффективно функционировать. Вы читаете эту статью ради экзамена или из интереса? В зависимости от цели меняются стратегии чтения и критерии контроля. Без цели вы не знаете, «до какой степени это достаточно». А оценка того, «достаточно ли этого», является одной из ключевых функций метапознания.

Стратегии. Когнитивные стратегии, выбранные для достижения цели. Столкнувшись с непонятным абзацем, вы заглядываете в словарь, пропускаете его или перечитываете несколько раз? Эти выборы зависят от ваших метакогнитивных знаний о задаче, о себе и об эффективности стратегий.

Люди с высоким уровнем метапознания могут гибко менять стратегии в зависимости от обратной связи; люди с низким уровнем, как правило, придерживаются одной стратегии, даже если она уже не работает.

Взаимодействие четырёх элементов: знания влияют на постановку целей и выбор стратегий, опыт в процессе работы даёт обратную связь и побуждает вас корректировать стратегии, а цели служат системой отсчёта для мониторинга. Суть метапознания заключается в том, чтобы «знать, что вы делаете со своими знаниями».

ИИ преуспевает в самом «познании»: сопоставлении шаблонов, интеграции информации, генерации речи. Но ИИ с трудом может по-настоящему осознавать, что именно он знает, понимать свои ограничения и активно корректировать стратегию в условиях неопределённости. Когда он генерирует ответ, он обычно не говорит: «В этом я не уверен» или «Эта часть выходит за пределы моих обучающих данных». Даже если и говорит, точность такого самоосознания намного уступает человеческой. Именно в этом заключается уникальная ценность человеческой метапознавательности.

1.3.6 Двухуровневая архитектура: базовая метапознавательность и приложения, основанные на метапознавательности

Сфера способностей, рассматриваемых в данной книге, весьма обширна: от осознания когнитивных границ до постановки вопросов, управления вниманием, ценностных

суждений и междисциплинарных связей. Чтобы избежать концептуального раздувания понятия «всё является метапознанием», необходимо различать два уровня.

Базовый уровень (метапознание в узком смысле, исходная модель Флавелла):

Осознание когнитивных границ: понимание того, что вы знаете, а чего не знаете (глава 3)

Когнитивный мониторинг: отслеживание собственных мыслительных процессов и состояний в режиме реального времени (главы 3, 6)

Когнитивная регуляция: гибкая корректировка когнитивных стратегий на основе результатов мони (главы 6, 9)

Прикладной уровень (высшие способности, обусловленные метапознанием):

Умение задавать точные вопросы: для этого требуется метапознание, чтобы определить, «правильно ли я задаю вопросы» (глава 4)

Управление вниманием: требует метапознания, чтобы осознать, «на чем я сосредоточен, а что игнорирую» (глава 5)

Оценка ценностей: требует метапознания для анализа «на чем основано мое суждение» (глава 7)

Распознавание и отбор междоменных связей: требуется метапознание, чтобы решить: «Стоит ли отслеживать эту, казалось бы, не имеющую отношения к делу мысль?» (глава 8)

Рис. 1-2: Двухуровневая архитектура — основные метакогнитивные навыки и их применение

Это разграничение имеет большое значение. Когда кто-то возражает, что «ИИ тоже может проводить междоменные аналогии» или «ИИ тоже учится корректировать неопределённость», он нападает на определённые функции прикладного уровня. Ядро — осознание, мониторинг и регулирование собственных когнитивных процессов — до сих пор остаётся самым слабым аспектом ИИ. Некоторые способности прикладного уровня могут постепенно переходить к ИИ, но пока ядро остаётся активным, вы сможете продолжать адаптироваться и осваивать новые способы применения.

Именно поэтому в этой книге метапознание сравнивается с операционной системой, а не с «рвом». Рв — это статическая защита, которая теряет эффективность, как только её прорвут. Операционная система — это динамическая инфраструктура, которую можно обновлять, адаптировать к новой среде и запускать на ней новые приложения. Вашей метапознавательной ОС не нужно всегда опережать ИИ — ей нужна способность к постоянной итерации.

1.4 Аргумент «от обратного»: признание сложности

Говорить о важности метапознания не означает, что другие аспекты можно игнорировать. Любая «теория человеческого превосходства», доведенная до крайности, приводит к искажению реальности, и есть несколько моментов, с которыми нужно честно признаться.

ИИ не заменил все ценности человека. Уход за больными, образование, творческое сотрудничество, реагирование на

кризисные ситуации — все эти виды деятельности по-прежнему требуют работы на месте, межличностного взаимодействия и физического присутствия. ИИ может составить диагностический отчет, но ему трудно взять за руку пациента, испытывающего тревогу; он может сгенерировать план урока, но ему трудно уловить недоумение в глазах учеников в классе и мгновенно скорректировать ход занятия. Настоящая книга посвящена работникам умственного труда и не охватывает всех трудящихся. Даже в рамках умственного труда степень влияния ИИ на различные должности различна.

Метапознание также не является панацеей. Некоторые должности по-прежнему в значительной степени зависят от накопленных знаний, например, анатомическая ориентация во время хирургической операции или цитирование положений закона в ходе судебных прений, когда требуется мгновенное воспроизведение профессиональных деталей. Метапознание помогает лучше применять знания и распознавать границы, но не может заменить накопление отраслевых знаний. Человек, четко осознающий границы своего познания, но не обладающий отраслевыми знаниями, по-прежнему не сможет выполнять работу, требующую глубоких профессиональных суждений. Оба аспекта взаимодополняют друг друга: метапознание — это «как использовать», а знания — «что использовать». (О структурных отношениях между познанием и метапознанием, а также о том, сможет ли ИИ развить

метапознание, подробно пойдет речь в следующей главе.)

Ключ заключается в смещении относительного преимущества. Раньше относительное преимущество человека в основном заключалось в том, «что он знает». Сегодня это преимущество во всё большем числе областей подвергается эрозии со стороны ИИ. Относительное преимущество смещается в сторону «как думать», «как судить» и «как корректировать». Знания по-прежнему важны, но в условиях, когда ИИ может легко восполнить их недостаток, метакогнитивные способности становятся ключевым критерием, позволяющим отличить «заменимое» от «трудно заменимого». Вам не нужно отказываться от обучения, но необходимо перераспределить внимание: перейти от «накопления большего количества знаний» к «лучшему осознанию и применению уже имеющихся знаний».

Признание этих сложностей помогает нам избежать дихотомии: не впадать ни в чрезмерный пессимизм (человечество будет полностью заменено), ни в чрезмерный оптимизм (метакогнитивные способности решат все проблемы). Нужно трезво осознавать тенденции и находить в них свою точку опоры.

1.5 Мегкогнитивные моменты: ваши упражнения на рефлексию

Теория обретает жизнь только тогда, когда она воплощается в личном опыте. Два приведенных ниже упражнения не требуют длинных ответов — главное, честно отнестись к се-

бе. Цель не в том, чтобы «сделать правильно», а в том, чтобы пробудить осознанность и заставить вас обратить внимание на те когнитивные процессы, которые обычно упускаются из виду.

Упражнение 1: Разберите пять дел, которые вы сделали сегодня

Вспомните пять конкретных дел, которые вы сделали сегодня (или вчера) на работе. Это может быть написание письма, участие в совещании, заполнение таблицы, анализ набора данных, принятие решения — любая задача, требующая когнитивного участия. Постарайтесь выбрать дела разного характера, чтобы сравнение было более наглядным.

По каждому делу задайте себе три вопроса:

Может ли ИИ выполнить это дело? Если да, то в какой степени? Может ли он полностью заменить вас или только помочь? Попытайтесь конкретизировать: какую часть может выполнить ИИ, а какую — нет?

Какую часть ИИ не может выполнить? Понимание контекста? Определение приоритетов? Управление межличностными отношениями? Принятие окончательной ответственности? Этот вопрос часто заставляет вас провести разграничение между «автоматизируемой частью задачи» и «неавтоматизируемой частью».

Почему ИИ не может выполнять эту часть? Попытайтесь объяснить это одним предложением — это обычно помогает выявить «уникально человеческий» аспект данного дела.

Требуется ли физическое присутствие? Необходима ли моральная ответственность? Нужно ли принимать решения в условиях неопределённости? Требуется ли понимание невысказанного подтекста?

Многие обнаруживают: ИИ, как правило, способен выполнять задачи и генерировать результаты; то, что ИИ не может сделать, — это суждения, выбор, ответственность и принятие решений в условиях неопределённости. За последним стоит именно метапознание. Если вам кажется, что ИИ способен выполнить какую-то задачу полностью, задайте себе ещё несколько вопросов: кто решает, делать это или нет? Какой уровень выполнения считается достаточным? Кто несёт ответственность за результат? Ответы на эти вопросы часто указывают на роль метапознания человека.

Упражнение 2: Когда вы в последний раз изменили своё мнение по какому-то важному вопросу?

Задумайтесь: когда в последний раз вы действительно изменили своё мнение о каком-то важном вопросе под влиянием новых доказательств или аргументов? О чём шла речь? Что стало причиной вашего изменения?

Многим будет сложно ответить на этот вопрос. Не потому, что они никогда не меняли своего мнения, а потому, что изменения обычно происходят постепенно и неосознанно, и мы редко чётко отмечаем «тот момент, когда я передумал». Возможно, однажды мы вдруг обнаружим: «Похоже, я больше так не думаю», но не сможем точно сказать, какой кон-

кретный момент или какие конкретные доказательства привели к этой перемене.

Одним из важных аспектов метапознания является осознание обновления убеждений. Осознание того, что «раньше я так думал, но теперь у меня появились новые основания, поэтому я должен скорректировать своё мнение», — само по себе является проявлением метапознания в действии. Это означает, что у вас не только есть точка зрения, но вы также способны наблюдать за ней и активно обновлять её при изменении доказательств. Если вы можете чётко вспомнить такой момент, это означает, что в тот момент ваше метапознание было активным. Если вам трудно вспомнить такой момент, возможно, стоит более активно развивать эту способность в повседневной жизни. Например, регулярно спрашивайте себя: «Какие важные взгляды у меня изменились за последний месяц? Почему?»

1.6 Подведение итогов и следующие шаги

Основная идея этой главы: происходит структурный переворот в асимметрии знаний. От шахмат до письма и анализа — преимущество человека в том, «что он знает», постепенно уступает место машинам. В условиях этого переворота ориентир ценности человека смещается: от того, что вы знаете, к тому, как вы осмысливаете то, что знаете. Это и есть метапознание.

Стоит запомнить несколько ключевых моментов:

Переворот уже произошел. Во все большем числе отрас-

лей утверждение «человек знает больше, чем машина» больше не верно.

Метапознание — это модернизируемая операционная система, а не статичный защитный вал. Способность осознавать, контролировать и корректировать собственные когнитивные процессы — это когнитивная инфраструктура, в которую сегодня наиболее целесообразно инвестировать.

Исследование Шариати и др. (2025) содержит предупреждение: ИИ делает человека умнее, но, возможно, менее мудрым. Чем лучше человек умеет пользоваться ИИ, тем менее точна его самооценка.

Флавелл (1979) разделил метапознание на четыре компонента: знания, опыт, цели и стратегии. Они взаимодействуют друг с другом, образуя «познание о познании».

В то же время мы признаем сложность реальности: ИИ не заменил всего, а метапознание — не панацея; ключ заключается в направлении перераспределения относительных преимуществ.

Если эта глава заставила вас задуматься, вы уже сделали первый шаг. Но прежде чем перейти к конкретным метакогнитивным способностям, необходимо ответить на более фундаментальный вопрос: какова, собственно, связь между когнитивными процессами и метакогнитивными процессами? Почему считается, что это две разные уровни умственной деятельности? Почему ИИ силен на когнитивном уровне, но слаб на метакогнитивном? В следующей главе будет

заложена структурная основа, проходящая красной нитью через всю книгу.

Список литературы

Shariati, M. и др. (2025). «AI Makes You Smarter but None the Wiser: The Disconnect Between Performance and Metacognition». *Computers in Human Behavior*. DOI: 10.1016/j.chb.2025.108662 — Основной вывод: использование ИИ повышает производительность, но ухудшает метакогнитивную калибровку; чем выше уровень ИИ-грамотности, тем менее точна самооценка.

Флавелл, Дж. Х. (1979). «Метапознание и когнитивный мониторинг: новая область исследований в области когнитивно-развивающей психологии». *American Psychologist*, 34(10), 906–911. — Основополагающая статья по концепции метапознания, в которой определены четыре компонента метапознания: знания, опыт, цели и стратегии.

Brynjolfsson, E., & McAfee, A. (2014). *«Вторая эра машин: работа, прогресс и процветание в эпоху блестящих технологий»*. W. W. Norton. — Структурные изменения в соотношении технологий и человеческих способностей, служащие макронарративной основой для концепции «переворота».

Агравал, А., Ганс, Дж., и Голдфарб, А. (2018). *«Машины предсказания: простая экономика искусственного интеллекта»*. Harvard Business Review Press. ISBN: 978-1633695672 — Как ИИ в качестве «машин предсказа-

ния» меняет распределение полномочий при принятии решений, предлагая экономическую перспективу для трансформации роли работников умственного труда.

Nelson, T. O., & Narens, L. (1990). «Metamemory: A Theoretical Framework and New Findings.» *Psychology of Learning and Motivation*, 26, 125–173. — Двухуровневая модель метапамяти, включающая мониторинг и контроль, обеспечивает теоретическую основу для понимания механизмов функционирования метапознания.

Глава 3: Знать, чего ты не знаешь / Knowing Your Unknowns

Ключевой вопрос: способность, которой ИИ лишен больше всего — осознание границ собственных знаний / The capability AI lacks most: perceiving the boundaries of your own knowledge

3.1 Вступительная история: «обещание» чат-бота

В ноябре 2022 года умерла бабушка Джейка Моффатта, жителя Ванкувера. Он зашёл на официальный сайт авиакомпании Air Canada, чтобы забронировать билет домой. В правом нижнем углу сайта появилось всплывающее окно чата — ИИ-бот службы поддержки авиакомпании.

Моффатт спросил: «Предусмотрены ли скидки на поездки в связи с траурными обстоятельствами?»

Бот ответил четко сформулированным и логически выстроенным объяснением: пассажир может сначала приобрести билет по полной стоимости, а затем в течение 90 дней

после поездки подать заявку на скидку в связи с похоронами, после чего ему будет возвращена часть стоимости билета. Тон ответа был профессиональным, условия — четкими, и даже были заботливо перечислены шаги для подачи заявки. Это читалось совершенно как официальная политика компании.

Моффатт поступил именно так. Он купил билет по полной стоимости, полетел домой на похороны, а затем в установленный срок подал заявку на возмещение.

«Эйр Канада» отказала. Причина: согласно политике компании в отношении скидки в связи со смертью близкого человека, заявку необходимо подавать до вылета, а подача заявки задним числом не допускается. Процедура, о которой рассказал Моффатту бот, была ошибочной от начала до конца. Такой «политики» вообще не существовало.

Моффатт подал иск против авиакомпании в Канадский суд по гражданскому арбитражу. Стратегия защиты авиакомпании оказалась удивительной — они утверждали, что чат-бот является «отдельным юридическим лицом» (a separate legal entity), его высказывания не отражают позицию компании, и компания не несет ответственности за то, что говорит бот.

Арбитр Кристофер Риверс (Christopher Rivers) в своем решении от февраля 2024 года отклонил этот довод защиты. В решении говорится: «Air Canada не объяснила, почему пассажир должен самостоятельно проверять, соответствует ли

информация, предоставленная чат-ботом, данным на других страницах официального сайта. Если вы размещаете ИИ на своем сайте для обслуживания клиентов, вы обязаны нести ответственность за его результаты». Моффатт получил компенсацию.

Сумма компенсации по этому делу невелика, но оно прекрасно иллюстрирует ситуацию, когда все три стороны «не знают, чего они не знают»:

ИИ не знает, чего он не знает. У чат-бота нет никаких встроенных механизмов, позволяющих ему оценить точность выдаваемой им информации о правилах. Он не колеблется, не помещает пометку «Эта информация не проверена» и не предлагает в случае неопределённости обратиться к оператору службы поддержки для подтверждения. Он с полной уверенностью выдал совершенно неверную информацию.

Пассажир не знал, чего он не знает. У Моффатта были все основания доверять этому ответу — он появился на официальном сайте авиакомпании, был сформулирован профессионально и логически последовательно. У него не было причин подозревать, что крупная авиакомпания разместит в своей системе обслуживания клиентов ненадежный источник информации. Он не знал, что общается с языковой моделью, не способной проверять факты.

Авиакомпания не знала, чего она не знает. «Эйр Канада» внедрила этот чат-бот, но, по-видимому, не создала никаких

механизмов для контроля соответствия его ответов фактической политике компании. Только после получения решения арбитражного суда они обнаружили, что их ИИ все это время «обещал» клиентам несуществующие правила.

У провалов всех трех сторон есть одна общая причина: никто — и ни одна система — не проверил информацию на уровне «действительно ли мы знаем, что эта информация верна». Именно об этом ключевом умении и пойдет речь в данной главе: осознание когнитивных границ — понимание того, что мы знаем, чего не знаем, а также, что еще опаснее, осознание того, о каких «незнаниях» мы даже не подозреваем.

Реальные последствия: дело Air Canada — не единичный случай. В 2023 году нью-йоркский адвокат Стивен Шварц (Steven Schwartz) представил в деле *Mata v. Avianca Inc.* шесть вымышленных судебных прецедентов, сфабрикованных ChatGPT, за что был подвергнут санкциям федеральным судьей — он также никогда не проверял достоверность результатов, выдаваемых ИИ. В сфере медицины в многочисленных отчетах фиксировались случаи, когда системы ИИ, отвечающие на вопросы, давали по поводу редких заболеваний рекомендации по лечению, которые выглядели профессионально, но были полностью ошибочными. От залов судебных заседаний до служб поддержки и врачебных кабинетов — схема удивительно одинакова: ИИ не осознает, что выдумывает, люди не осознают, что их вводят в заблужде-

ние, а организации, внедряющие ИИ, не осознают, что у них отсутствует надзор.

3.2 Основная теза: опасность «незнания о собственном незнании»

Флавелл (Flavell, 1979) в своей работе, заложившей основу концепции метапознания (metacognition), отмечает: когнитивный мониторинг (cognitive monitoring), то есть осознание и оценка собственных когнитивных процессов и состояний, является ключевой способностью, отличающей человека от многих систем обработки информации. Самый опасный когнитивный пробел заключается в том, что вы не знаете, чего вы не знаете.

Галлюцинации ИИ, то есть генерация казалось бы разумного, но полностью вымышленного контента, по сути являются машинной версией того же самого провала. ИИ не знает, чего он не знает. Он будет уверенно придумывать цитаты, фабриковать данные и выдумывать события, не имея под собой реальных знаний. Это происходит из-за отсутствия у него осознания границ собственных знаний.

Если человек развит мощную метакогнитивную калибровку (metacognitive calibration), то есть способность соотносить уровень своей уверенности с фактической точностью, он сможет улавливать ошибки, упускаемые ИИ и людьми, чрезмерно полагающимися на ИИ. Ключевой способностью является более острое, чем у ИИ, осознание того, что «на самом деле я этого не знаю», а объем знаний при этом имеет

второстепенное значение.

Рис. 3-1: Четыре уровня когнитивных границ

3.3 Доказательства и мыслительные эксперименты

3.3.1 Иллюзии ИИ: окно для наблюдения за отсутствием метапознания

Иллюзии ИИ предоставляют нам уникальную возможность понять последствия «отсутствия метапознания». В период с 2023 по 2025 год множество пользователей сообщали о случаях, когда модели большого языкового объёма (LLM) придумывали ссылки, выдумывали научные статьи и фальсифицировали статистические данные. Например, более половины ссылок на судебные прецеденты, сгенерированных одним из инструментов для юридических исследований, не существовало в базе данных; юристы обнаружили это только тогда, когда сослались на них в суде, и выяснилось, что в судебных протоколах этих дел вообще не было. Одна из систем ответов на медицинские вопросы давала по поводу редких заболеваний рекомендации по приему лекарств, которые казались профессиональными, но на самом деле были полностью ошибочными, что едва не привело к задержке лечения пациента. Еще один ученый обнаружил, что в обзоре литературы, написанном для него ИИ, названия, авторы и журналы многих «ключевых ссылок» были полностью вымышленными, хотя читались как настоящие.

Ещё более тревожно то, что эти «галлюцинации» часто появляются именно в самых «бегло» сформулированных от-

ветах ИИ. Столкнувшись с неопределённым вопросом, он ни за что не признает: «Я не знаю». Вы всегда получаете ответ, который выглядит полным, связным и подкреплённым ссылками. Разрыв между беглостью изложения и точностью — типичный симптом отсутствия метапознания.

У этих систем есть одна общая черта: они не «знают», что они не знают. У них отсутствует внутренний механизм метакогнитивного мониторинга (metacognitive monitoring) для оценки надежности вывода. Они выдают правильные ответы и полную чепуху с одинаковой уверенностью, и для человеческого пользователя на первый взгляд их невозможно отличить.

Если человек утратит способность калибровать собственную уверенность, то, полагаясь на ИИ, он будет повторять одни и те же ошибки: принимать уверенность ИИ за свою собственную и считать границы ИИ границами знания.

Конец ознакомительного фрагмента.