12+

АЛЕКСАНДР НОВИКОВ-АНДРИЕНКО

COGNITIVE HARNESS

Архитектура человеческого мышления
в эпоху агентного искусственного интеллекта

Александр Новиков-Андриенко

Cognitive Harness. Архитектура организации человеческого мышления в эпоху агентного искусственного интеллекта

https://litres.ru/74320189
ISBN 9785007093170

Аннотация

Что, если архитектура AI-агентов способна помочь человеку лучше понять собственное мышление? «Cognitive Harness» предлагает авторскую модель

организации внимания, целей, памяти, решений и обратной связи. Книга исследует сознание, ум, эго и взаимодействие человека с искусственным

интеллектом, а также показывает, как сохранить субъектность и ответственность в наступающей агентской экономике.

Содержание

Cognitive Harness 7
- Предисловие 8
- Как читать эту книгу 10
- От ответа к действию 13
- Второе открытие 18
- Что именно показал масштаб 21
- Надежность длинного цикла 22
- Сильнейшее возражение 23
- Самая опасная ошибка 25
- Три вопроса вместо одного 27
- Что действительно общее 30
- Что не является аналогией 32
- Тело, развитие и отношения 34
- Четыре теста хорошего переноса 36
- Аналогия как исследовательский инструмент 38
- Практическая цена ошибки 39
- Методологический контракт 40
- Эксперимент с двумя системами 42
- Интеллект, агентность и автономность 44
- Почему одной модели недостаточно 45
- Семейство агентных архитектур 46
- Memory: состояние, а не склад 48
- Evaluator: проблема критерия 49

Reflection: полезная и опасная рекурсия 50
Наблюдаемость и журнал решения 52
Когда агентность избыточна 53
Архитектура отказов 54
Что изменилось для человека 56
Предварительное определение 62
Чем Harness не является 63
КУЛЬТУРА И СОЦИАЛЬНАЯ СРЕДА 66
↕ 67
ЦЕННОСТИ И ЗНАЧИМОСТЬ ↔ КООРДИНАЦИЯ ↔ ДЕЙСТВИЕ В СРЕДЕ 68
Шесть задач первой карты 69
Время как несущая ось 72
Внутренний и внешний контур 74
Порядок дальнейшего исследования 76
Свет в комнате и доступ к выключателю 78
Феноменальный опыт и доступ 80
Внимание не равно сознанию 82
Global Workspace: глобальная доступность 84
Predictive processing: ожидание и доступ 87
Сознание в первой карте Harness 89
Сильнейшее возражение 90
Задача без недостатка интеллекта 92
Аффект: оценка до аргумента 94
Любопытство и внутренняя мотивация 96
Интерес, цель и ценность 99

Социальное производство интереса 100
Практический случай 101
Сильнейшее возражение 102
Восприятие начинается до сигнала 104
От пассивного восприятия к конструктивному 105
Predictive processing 106
Воплощенное предсказание 109
Неопределенность нескольких типов 110
Мозг не заканчивается ответом 111
LLM как демонстрация, а не модель мозга 112
Практический случай: ранняя фиксация 113
Сильнейшее возражение 114
Чья рука движется? 116
Почему термин опасен 117
Минимальная самость 118
Интероцепция: модель внутреннего тела 119
Нарративная самость 120
Агентность и авторство 121
Социальная идентичность 122
Фрейд: историческая модель конфликта 123
Нейронаука: распределенные процессы 125
Конец ознакомительного фрагмента. 128

Cognitive Harness
Архитектура организации человеческого мышления в эпоху агентного искусственного интеллекта

Александр Владимирович Новиков-Андриенко

© Александр Владимирович Новиков-Андриенко, 2026

ISBN 978-5-0070-9317-0
Создано в интеллектуальной издательской системе Ridero

Cognitive Harness

Архитектура организации человеческого мышления в эпоху агентного искусственного интеллекта

Александр Новиков-Андриенко
2026
Table of Contents
Cognitive Harness

Архитектура человеческого мышления в эпоху агентного искусственного интеллекта

Автор: Александр Новиков-Андриенко

Редакция: 1 августа 2026 года

Эта книга предлагает концептуальную модель и исследовательскую программу. Она не заменяет медицинскую, психологическую или иную профессиональную помощь. Аналогии между человеческим мышлением и AI-архитектурами не означают биологического тождества человека и машины.

Предисловие

Человечество долго пыталось понять разум, изучая мозг, поведение и субъективный опыт. Развитие искусственного интеллекта добавило к этим перспективам ещё одну — инженерную. Стало видно, что качество интеллектуальной системы определяется не только вычислительной мощностью. Не менее важны постановка цели, распределение внимания, доступ к памяти, выбор инструментов, проверка результата, право действовать и способность изменить стратегию.

Эта книга начинается с рабочей гипотезы: архитектуры современных агентных систем могут дать полезный язык для разговора об организации человеческого мышления. Не потому, что мозг является компьютером, сознание — интерфейсом, а личность — программой. И не потому, что инженерная схема способна исчерпать человеческий опыт. Ценность аналогии уже и скромнее: она позволяет различить интеллект и управление интеллектом, действие и право на действие, накопление опыта и обновление модели мира.

Название *Cognitive Harness* обозначает не отдельный участок мозга и не внутреннего управляющего человечка. Это имя для распределённого контура, в котором цели, ограничения, внимание, память, планирование, действие и обратная связь согласуются во времени. Его можно искать внутри индивидуальной практики, во внешних инструментах, в ко-

манде и в гибридной системе «человек—AI».

Читатель встретит в тексте пять типов утверждений: данные, интерпретации, рабочие аналогии, авторские гипотезы и сценарии. Они намеренно разделены. Там, где наука даёт несколько конкурирующих объяснений, книга не будет превращать одно из них в установленный факт. Там, где предлагаемая модель выходит за пределы данных, это будет названо гипотезой. Главная задача — не объявить новую окончательную теорию мышления, а построить достаточно точный язык, чтобы эту модель можно было обсуждать, применять, критиковать и проверять.

Как читать эту книгу

Первая часть показывает, почему агентный AI заставил по-новому поставить старый вопрос об организации интеллекта. Вторая разбирает человеческую сторону аналогии: сознание, интерес, прогнозирующий мозг и модель себя. Третья формулирует Cognitive Harness как архитектуру и исследует её основные контуры. Четвёртая превращает модель в практические протоколы и рассматривает жизнь в агентской экономике. Пятая задаёт программу эмпирической проверки и прямо перечисляет условия, при которых модель окажется слабой или лишней.

Практические упражнения не являются валидированным диагностическим тестом. Их задача — сделать архитектуру наблюдаемой в собственной работе: увидеть дрейф цели, перегрузку внимания, плохую память решения, чрезмерное делегирование или отсутствие обратной связи. Полезность упражнения сама по себе ещё не доказывает теорию. Именно поэтому книга заканчивается не манифестом, а исследовательской программой и сильной критикой собственных оснований.

Глава 1. Мы случайно построили зеркало собственного мышления

Иногда инженерия оказывается быстрее философии

В конце 2022 года миллионы людей открыли окно чата и

впервые столкнулись не с поисковой строкой, а с системой, способной продолжить почти любую интеллектуальную ситуацию. Она объясняла понятия, писала код, меняла стиль текста, предлагала планы и поддерживала диалог. Ошибки были заметны с первых дней: вымышленные ссылки, уверенные неточности, неспособность отличить правдоподобие от факта. Но даже с учетом этих ограничений событие оказалось больше выпуском нового продукта.

Большая языковая модель сделала наблюдаемой способность, которую прежде было легко связывать только с целостным человеческим субъектом: порождение связного ответа в открытом контексте. Она не решила проблему сознания и не доказала, что машина понимает слова так же, как человек. Она предоставила инженерный контрпример более скромной, но важной интуиции: сложное языковое поведение может возникать из статистического обучения на больших массивах данных без заранее записанного каталога правил для каждого вопроса.

История этой технологии началась задолго до появления массовых чат-интерфейсов. Нейронные языковые модели учились предсказывать следующий элемент последовательности; архитектура Transformer позволила эффективнее работать с отношениями между элементами контекста; масштабирование данных, вычислений и параметров расширило диапазон задач; обучение следованию инструкциям и обратная связь от людей сделали взаимодействие удобнее. Каж-

дый этап имел собственные научные и инженерные основания. Поэтому соблазнительно рассказать эту историю как прямую лестницу к «машинному разуму», но такая ретроспектива скрывает случайности, тупики и нерешенные проблемы.

С философской точки зрения важнее другое. Исследователи строили систему обработки языка, а получили новый инструмент для исследования наших критериев мышления. Если связный текст, объяснение и план уже недостаточны, чтобы уверенно приписать сознание, что именно мы считали признаком разума? Если модель может создать хороший вариант решения и все же не способна самостоятельно определить, какую проблему следует решать, где проходит граница между интеллектом и агентностью?

Факт / данные. Современные языковые модели обучаются на задаче предсказания и последующей настройке поведения; их уверенный текст не является надежным индикатором фактической истинности или субъективного понимания.

От ответа к действию

Первое поколение массового опыта с LLM было сосредоточено на ответе. Пользователь формулировал запрос, модель возвращала текст. В этой схеме человек удерживал почти всю архитектуру задачи: выбирал цель, собирал контекст, решал, когда обратиться к модели, проверял результат и переносил его в мир.

Если модель писала письмо, человек выбирал адресата и нажимал «отправить». Если предлагала код, человек запускал его и отвечал за последствия. Если перечисляла варианты лечения, медицинское решение оставалось вне системы. Модель могла быть сильным генератором, но цикл действия замыкался человеком.

Затем вокруг моделей начали появляться дополнительные компоненты. Системе давали доступ к поиску, калькулятору, среде исполнения кода и корпоративным данным. Добавляли память о предыдущих шагах, планировщик, проверки результата и правила повторной попытки. Она переставала быть только собеседником и становилась частью агентной системы.

Агент здесь не означает сознательное цифровое существо. Это операционное название системы, которая поддерживает цикл: получает задачу, наблюдает доступное состояние, выбирает следующий шаг, использует разрешенные инструмен-

ты, оценивает результат и продолжает до выполнения условия остановки. Степень автономности может быть небольшой: от подготовки черновика до ограниченного выполнения многошагового процесса.

Этот переход обнаружил странную вещь. Повышение качества базовой модели не решало автоматически проблемы действия. Сильная модель могла выбрать неправильный инструмент, потерять цель в длинной цепочке, повторить неудачный шаг, принять данные из недоверенного источника за инструкцию или не понять, когда следует остановиться. Надежность зависела от системы вокруг модели.

Инженерам пришлось проектировать цели, память, маршрутизацию, права доступа, наблюдаемость, оценку и эскалацию человеку. В центре внимания оказался не только интеллект компонента, но и организация его работы.

Вопрос, который никто не собирался задавать

История AI долго вращалась вокруг сравнения способностей. Может ли машина играть, распознавать изображение, переводить, доказывать теорему, поддерживать разговор? Такой способ измерения естественен: он позволяет выделить задачу и сравнить результат.

Агентные системы изменили вопрос. Даже если компонент способен решить каждую отдельную задачу, кто определяет их последовательность? Кто решает, что наблюдать, какие данные считать достаточными, какой риск приемлем и что делать после ошибки? Эти вопросы не сводятся к допол-

нительной порции интеллекта. Они описывают архитектуру целенаправленного поведения.

В программной системе ответы воплощаются в коде, интерфейсе и организационных правилах. Цель формулирует пользователь или другой процесс. Доступ ограничивает политика безопасности. Память хранит состояние. Evaluator сравнивает результат с тестом или критерием. Лимит шагов останавливает цикл. Человек вмешивается при исключении.

Но кто выполняет сопоставимые функции в человеческом мышлении? Обычный ответ — «сам человек». Он верен на уровне повседневной ответственности и почти ничего не объясняет на уровне механизма. Человек не является простым центральным оператором собственного мозга. Намерения конкурируют, внимание захватывается, память реконструирует прошлое, привычки запускают действия раньше осознанного решения. Тем не менее связное поведение часто сохраняется месяцами и годами.

AI не дает готового объяснения психики. Он позволяет точнее сформулировать проблему: каким образом множество разнородных процессов образуют контур, способный удерживать направление, замечать ошибку и менять способ действия?

Рабочая аналогия. Модель внутри AI-агента похожа не на всего человека, а на один мощный компонент более широкой системы. Это сравнение помогает различить способность генерировать варианты и архитектуру, которая превра-

щает варианты в длительное действие.

Почему язык ввел нас в заблуждение

Связная речь обладает сильным социальным эффектом. В человеческом мире она обычно исходит от существа с телом, биографией, интересами и ответственностью. Когда похожая форма возникает в машине, мы автоматически достраиваем привычного собеседника. Грамматика усиливает иллюзию: система говорит «я думаю», «я помню», «я решил».

Антропоморфизм не является просто глупостью пользователя. Это эффективная стратегия социального мозга: воспринимать поведение через намерения. Но в работе с AI она нуждается в корректировке. Первое лицо в ответе модели не подтверждает наличие устойчивого «я»; объяснение собственного вывода может быть правдоподобной реконструкцией, а не доступом к причинной истории вычисления.

Существует и обратная ошибка — описывать человека как несовершенную машину. Если LLM порождает текст через вероятностное продолжение, нельзя заключить, что человеческая речь является тем же процессом. Человек воплощен, развивается в отношениях, действует из потребностей, испытывает боль и несет последствия. Сходство результата или абстрактной функции не устанавливает тождество реализации.

Эта книга будет удерживать обе границы. Она не очеловечивает AI и не механизирует человека. Инженерная система используется как зеркало: оно показывает выбранную струк-

туру, но создано из другого материала и оставляет часть предмета вне отражения.

Второе открытие

Публичное внимание было приковано к росту возможностей моделей. Однако для темы этой книги важнее открытие, сделанное при попытке превратить модель в надежного агента: качество поведения определяется контуром управления не меньше, чем качеством генерации.

Представим две системы на одной базовой модели. Первая получает расплывчатую инструкцию, широкий доступ к инструментам и завершает работу по собственному текстовому суждению. Вторая уточняет цель, использует ограниченные права, хранит состояние, проверяет ключевые результаты независимым способом, ведет журнал и передает человеку необратимые решения. Разница между ними не находится «в интеллекте» модели. Она возникает в архитектуре.

То же наблюдение можно сделать о человеке. Два специалиста сходных способностей получают разные результаты из-за того, как формулируют задачу, защищают внимание, используют записи, запрашивают обратную связь и пересматривают предположения. Это не означает существование особого внутреннего устройства. Речь идет о конфигурации процессов и внешних опор.

Именно эту конфигурацию книга предлагает назвать Cognitive Harness. Сейчас достаточно предварительной ин-

туиции: это контур, который связывает способность думать с направлением, действием и обучением. Полное определение потребует пройти путь через сознание, интерес, генерацию гипотез и модель себя, а затем снять проблему внутреннего управляющего.

От персонального помощника к агентной экономике

Когда AI способен не только советовать, но и выполнять ограниченные действия, меняется экономическая роль программного обеспечения. Появляется среда, в которой агенты ищут информацию, сравнивают предложения, вызывают цифровые услуги, координируют расписания и готовят транзакции от имени людей и организаций. В этой книге такая формирующаяся среда называется агентской экономикой.

Речь не идет о завершенном мире автономных машин. Современные системы хрупки, зависят от инфраструктуры и человеческих полномочий. Но направление уже меняет вопрос о человеческом преимуществе. Если производство черновиков, вариантов и стандартных операций дешевеет, ценность смещается к выбору проблемы, постановке границ, проверке и ответственности.

Один человек может координировать несколько специализированных агентов. Это расширяет способность действовать и одновременно увеличивает поверхность ошибки. Неправильная цель теперь масштабируется быстрее. Невнимательно выданное полномочие превращает текстовую неточность во внешний эффект. Память помощника

удобна, но создает риск приватности и фиксации устаревшего образа пользователя.

Поэтому наступление агентской экономики усиливает, а не отменяет человеческий контур. Главным навыком становится не умение вручную выполнять каждый шаг, а способность сохранять авторство цели, распределять доверие и понимать момент, когда действие должно вернуться к человеку.

Что именно показал масштаб

Развитие больших моделей часто описывают словом «масштабирование», как будто увеличение параметров и данных объясняет все появившиеся способности. Наблюдение тоньше: при изменении масштаба и обучения одна архитектура успешнее выполняет широкий набор задач, но границы зависят от теста, формулировки и инструментов.

Для книги важен инженерный результат: общая модель стала переносимым компонентом. Вместо отдельной программы для каждого узкого действия появился генератор, который можно включать в разные контуры через язык и инструменты. Часть вариативности поручается модели, а жесткими остаются границы, тесты и полномочия.

Сходство с организацией человеческой работы полезно и здесь же заканчивается. Человек не был предварительно обучен как универсальная модель, которую затем подключили к телу. Его способности формировались в действии, отношениях и развитии.

Надежность длинного цикла

Отдельный ответ может выглядеть надежным, но многошаговая задача требует серии успешных переходов. Ошибка раннего шага меняет контекст следующих. Поэтому впечатляющая демонстрация не гарантирует автономную работу в течение дня.

Надежность строится иначе, чем красота ответа. Система ограничивает область действия, проверяет промежуточное состояние, делает операции обратимыми и передает исключения. Формальный тест ловит формат и арифметику, но хуже — неуместную цель и социальный ущерб. Evaluator должен быть многослойным: машинная проверка для формального правила, внешний источник для факта и человеческое суждение для ответственности.

Высокий человеческий интеллект также не гарантирует надежности длинного проекта. Усталость, смена контекста и задержанная обратная связь создают цепочки отклонений. Контрольные точки нужны не потому, что человек «как AI», а потому, что ограниченные агенты сталкиваются с общей проблемой длительности.

Сильнейшее возражение

Можно возразить, что книга просто переносит модную инженерную лексику на старые проблемы психологии. Внимание, исполнительные функции, метакогниция и саморегуляция изучаются десятилетиями. Зачем добавлять еще одно слово?

Это возражение задает стандарт, которому модель обязана соответствовать. Новый термин оправдан только в том случае, если он связывает процессы на полезном уровне, обнаруживает классы сбоев, которые иначе остаются разрозненными, и предлагает проверяемые вмешательства. Если Cognitive Harness окажется лишь переименованием самоконтроля, от него следует отказаться.

Пока его потенциальный вклад состоит в трех акцентах: координации внутреннего и внешнего; полном цикле от выбора цели до обновления после действия; сравнении человеческой агентности с инженерными системами без утверждения их тождества. Поздние главы проверят, достаточно ли этого для самостоятельной исследовательской программы.

Следующий шаг — установить правила сравнения. Прежде чем использовать AI как зеркало, необходимо точно понять, что в нем отражается, а что принадлежит только поверхности зеркала.

Факт / данные. Языковая способность, фактическая надежность и автономное действие являются различимыми характеристиками AI-систем; агентное поведение зависит от внешней архитектуры управления и инструментов.

Гипотеза Cognitive Harness. Инженерный опыт создания AI-агентов может дать полезную рамку для исследования координации человеческого мышления.

Проверка. Добавляет ли понятие Cognitive Harness объяснительную и практическую ценность сверх существующих теорий контроля и саморегуляции.

Источники главы

— Vaswani, A. et al. «Attention Is All You Need».

— Brown, T. B. et al. «Language Models are Few-Shot Learners».

— Ouyang, L. et al. «Training Language Models to Follow Instructions with Human Feedback».

— Bender, E. M. et al. «On the Dangers of Stochastic Parrots».

— Weizenbaum, J. *Computer Power and Human Reason*.

Глава 2. Что общего между человеком и агентом?

Самая опасная ошибка

Сильная аналогия меняет способ видеть предмет. Именно поэтому она опасна. Сказав, что память похожа на хранилище, мы начинаем искать в ней неизменные записи. Назвав мозг компьютером, ожидаем программу и центральный процессор. Сравнив человека с AI-агентом, легко решить, будто внутренние процессы различаются только материалом.

Эта книга не делает такого вывода. Человеческий организм и программная система имеют разную историю возникновения, физическую реализацию и связь со средой. Человек воплощен, уязвим, развивается в отношениях, обладает аффектом и биографической непрерывностью. AI-агент создается из моделей, кода, данных, интерфейсов и выданных полномочий. Даже если обе системы можно описать словами «цель», «память» и «действие», значения этих слов не совпадают автоматически.

Но отказ от тождества не делает сравнение бесполезным. Между буквальным равенством и поэтической метафорой существует функциональная аналогия: две системы могут решать сходную архитектурную задачу разными средствами. Крыло птицы и крыло самолета не одинаковы, однако сравнение подъемной силы продуктивно. Оно становится ошибочным, если из него вывести, что самолет должен махать крыльями или переживать полет.

Задача этой главы — заключить методологический контракт. Мы определим, на каком уровне сравнение работает, какие выводы допустимы и где аналогия должна остановиться.

Три вопроса вместо одного

Полезно различать три уровня анализа, предложенные Дэвидом Марром для исследования информационных систем. Первый спрашивает, какую задачу решает система и почему эта задача имеет смысл. Второй — какие представления и процедуры позволяют ее решить. Третий — как эти процедуры физически реализованы.

Это различение не универсально и не исчерпывает человеческое познание, но дисциплинирует аналогию. На функциональном уровне человек и агент могут сталкиваться с проблемой ограниченного внимания. На алгоритмическом уровне один использует биологически сформированные процессы, другой — программную маршрутизацию и выбор контекста. На физическом уровне нейронная ткань и вычислительная инфраструктура радикально различаются.

Большинство продуктивных сравнений книги относится к первому и частично второму уровням. Мы спрашиваем: как удерживается цель, как сохраняется состояние, как выбирается действие, как ошибка изменяет следующий цикл? Мы не утверждаем, что одинаковая диаграмма потоков доказывает одинаковый субъективный опыт или нейронный механизм.

Есть и четвертый уровень, особенно важный для человека: историко-социальный. Цели формируются не в пустоте.

Язык, нормы, институты и отношения определяют, что система замечает и считает допустимым. AI-агент тоже встроен в социальную среду через данные, владельца, интерфейс и политику доступа, но способ этой встроенности иной.

Карта сходств и различий

— Удержать направление. Человек: намерение, мотивация, обязательство; **AI-агент:** инструкция, objective, workflow; **Граница аналогии:** инструкция не равна переживаемому намерению

— Выбрать значимое. Человек: внимание, аффект, интерес; **AI-агент:** фильтр, routing, context selection; **Граница аналогии:** программный приоритет не доказывает интерес

— Сохранить прошлое. Человек: реконструктивная память, навыки; **AI-агент:** журнал, база, параметры, retrieval; **Граница аналогии:** машинная запись не равна воспоминанию

— Создать варианты. Человек: прогнозирование, воображение; **AI-агент:** генерация, поиск, планировщик; **Граница аналогии:** сходный выход не означает сходного механизма

— Действовать. Человек: телесное и социальное действие; **AI-агент:** вызов инструмента через разрешения; **Граница аналогии:** агент зависит от внешней инфраструктуры

— Оценить результат. Человек: эмоции, нормы, суждение; **AI-агент:** тест, метрика, evaluator; **Граница анало-**

гии: метрика не обладает собственной ценностью

— **Измениться. Человек:** обучение, развитие, перестройка привычек; **AI-агент:** обновление состояния, памяти или модели; **Граница аналогии:** масштаб и способ пластичности различны

Таблица показывает, почему слова требуют уточнения. Когда говорится, что агент «помнит», это может означать доступ к сохраненной записи. Когда человек помнит, воспоминание зависит от контекста и может менять значение. Когда агент «ставит цель», обычно речь идет о полученном или вычисленном критерии. Человеческая цель связана с телом, идентичностью, отношениями и ответственностью.

Что действительно общее

Первое сходство — **ограниченность**. Ни человек, ни реальная программная система не обладают полным знанием, временем и вычислительным ресурсом. Они вынуждены отбирать контекст, принимать решения при неопределенности и останавливаться до исчерпания всех возможностей.

Второе — **зависимость от представления задачи**. Один и тот же мир допускает разные описания, а описание определяет доступные действия. Если проблема сформулирована как недостаток мотивации, будут искаться способы усилить усилие; если как конфликт целей — способы выбора; если как плохая среда — изменение условий.

Третье — **цикличность**. Действие меняет среду и создает обратную связь. Система, которая не обновляется после результата, повторяет прошлую политику. Сходство цикла позволяет сравнивать ошибки постановки цели, наблюдения, оценки и обучения.

Четвертое — **распределенность функций**. У человека нет единственного центра, содержащего восприятие, память, ценности и действие в готовом виде. AI-агент также состоит из разных компонентов и внешних сервисов. Связное поведение является результатом их координации.

Пятое — **роль внешней среды**. Человек думает с помощью языка, записей, инструментов и других людей. Агент

зависит от интерфейсов, данных и разрешений. Граница системы определяется не корпусом или файлом программы, а устойчивым контуром взаимодействия.

Рабочая аналогия. Общим является не материал и не внутренний опыт, а класс задач координации при ограниченных ресурсах.

Что не является аналогией

Книга не утверждает, что искусственная нейронная сеть работает как биологический мозг во всех значимых отношениях. Термин «нейрон» исторически вдохновлен биологией, но математический элемент модели не воспроизводит клеточную сложность, химию и развитие нервной системы.

Она не утверждает, что языковая модель обладает сознанием. Научного консенсуса о достаточных критериях сознания нет даже для ряда биологических случаев. Поведенческая убедительность не закрывает этот вопрос.

Она не сводит эмоции к ошибкам или шуму. Аффективные процессы участвуют в оценке значимости, мобилизации действия, социальном понимании и обучении. Машинный score может выполнять отдельную функцию приоритизации, но не дает основания приписать переживание.

Она не считает память человека базой данных. Человеческое воспоминание реконструктивно, связано с навыками, телом и идентичностью. Аналогия с retrieval полезна только для вопроса, какой контекст становится доступным сейчас.

Она не отождествляет сознание, интеллект, агентность и автономность. Интеллект относится к способности решать классы задач; агентность — к поддержанию цикла действия; автономность — к степени независимости от внешнего управления; сознание — к феноменальному опыту и/или

доступности содержания в зависимости от теории. Эти характеристики могут сочетаться по-разному.

Наконец, книга не утверждает моральное равенство человека и программного агента. Ответственность за проектирование, выдачу полномочий и использование современных систем остается у людей и институтов, если не показано иное.

Тело, развитие и отношения

Сильнейший контраргумент функциональной аналогии состоит в том, что человеческое мышление нельзя отделить от воплощенности. Мозг регулирует живое тело, которое должно сохранять температуру, энергию, безопасность и отношения. Потребности не поступают человеку как внешний prompt. Они возникают из устройства организма и истории развития.

Ребенок учится действовать в совместном внимании с другими людьми. Язык приходит не как готовый корпус данных, а как часть привязанности, игры, конфликта и культуры. Самость формируется в признании и сопротивлении. Эти процессы не являются дополнительными входами к универсальному вычислителю; они участвуют в создании самого пространства значений.

AI-агент также имеет среду и историю настройки, но его зависимость устроена иначе. Он не поддерживает живой организм, если специально не включен в такую систему. Его «награда» — инженерная величина, а не доказательство желания. Его отказ может быть функцией политики, а не переживаемой границы.

Поэтому Cognitive Harness не должен становиться теорией всего человека. Это модель организации задач и действий, которая обязана оставлять место для воплощенности, аф-

фекта, развития и культуры. Там, где эти измерения определяют явление, инженерная аналогия становится слабее.

Когда функция зависит от материала

Функционализм предполагает, что некоторые свойства можно описывать через роль независимо от реализации. В инженерии это мощный принцип: один алгоритм запускается на разном оборудовании, а регулятор может быть механическим или электронным. Но перенос на сознание и личность остается философски спорным.

Даже одна и та же функция может существенно меняться из-за материала. Биологическая память забывает и реконструирует; это не только дефект, но способ обобщения и поддержания идентичности. Тело ограничивает скорость, создает эмоции и придает решениям ставку. Смертность делает время не просто вычислительным бюджетом.

Поэтому корректная формула звучит так: **частичная функциональная сопоставимость возможна, но ее границы устанавливаются отдельно для каждого свойства**. Нельзя однажды объявить архитектуру независимой от материала и затем переносить все выводы.

Четыре теста хорошего переноса

Перед переносом понятия из AI в описание человека полезны четыре проверки. Первая — **функциональная точность**: решают ли процессы одну задачу? Машинная память хранит запись, человеческая может поддерживать идентичность; слово еще не подтверждает общую функцию.

Вторая — **каузальная полезность**: позволяет ли аналогия предсказать эффект вмешательства? Если проблема агента в контексте, релевантное извлечение должно помочь. Если человеческая проблема вызвана конфликтом мотивации, тот же прием не сработает.

Третья — **устойчивость к различиям реализации**. Расчет может быть сопоставим функционально. Боль нельзя свести к числовому сигналу только потому, что оба влияют на приоритет.

Четвертая — **этическая безопасность**. Метафора работника как агента помогает описать workflow и может оправдать контроль, игнорирующий достоинство. Очеловечивание AI делает интерфейс удобным и способно скрыть ответственность владельца.

Если перенос не проходит тест, его следует сузить. Такая дисциплина превращает аналогию в метод, способный показать и место своего провала.

Пример: память человека и память агента

Пусть AI-помощник хранит предпочтения пользователя: короткие утренние встречи, перелеты без пересадок, работа над книгой. Система извлекает записи и предлагает расписание. Функционально она использует прошлое для текущего действия.

У человека похожий выбор связан не только с фактом. Воспоминание о перелете несет тревогу, история книги — обязательство, а предпочтение может измениться после болезни. Значение прошлого реконструируется вместе с текущей самостью.

Аналогия retrieval помогает спросить, какой контекст активирован. Она перестает работать, когда записи приписывают переживание или считают человека обязанным сохранять прежний выбор. Более того, помощник влияет на идентичность: постоянно предлагая старое, он делает изменение менее вероятным.

Отсюда практический вопрос: должна ли память агента иметь срок давности и механизм «забыть меня таким»? Он возникает именно на границе сходства и различия.

Аналогия как исследовательский инструмент

Хорошая аналогия выполняет три задачи. Она делает заметной структуру, порождает вопрос и допускает проверку собственной границы. Если сравнение только украшает текст, научная ценность невелика.

В этой книге AI-агент позволяет увидеть разделение между генератором и контуром действия. Из этого возникает вопрос: можно ли улучшить человеческое решение, изменив архитектуру постановки цели, памяти и проверки, не изменяя базовую способность рассуждать? Проверяемое следствие: одинаковые участники должны показывать различный результат в зависимости от организации контура, особенно в длительных и неопределенных задачах.

Но аналогия может и провалиться. Если выделенные «модули» не дают новых предсказаний, если вмешательства полностью объясняются известной саморегуляцией или если архитектурный язык систематически игнорирует тело и социальную власть, модель следует сузить.

Такой способ использования аналогии отличается от утверждения «человек — агент». Человек может быть описан как агент в одном исследовательском контексте и как организм, личность, участник отношений или гражданин в другом. Ни одно описание не исчерпывает предмет.

Практическая цена ошибки

Граница аналогии важна не только для философии. Если работодатель считает человека заменяемым узлом workflow, он может игнорировать смысл, здоровье и неформальное знание. Если пользователь считает AI полноценным ответственным субъектом, он передает системе решение, для которого у нее нет морального или правового статуса. Если разработчик очеловечивает evaluator, метрика получает незаслуженное доверие.

В агентской экономике эти ошибки масштабируются. Программные агенты могут вести поиск, готовить контракты и инициировать операции. Чем убедительнее их язык, тем легче забыть, что цель, полномочия и ответственность были заданы людьми и организациями. Функциональная аналогия должна помогать проектировать контроль, а не размывать авторство.

Для человека полезен симметричный вывод. Нельзя относиться к себе как к плохо настроенному агенту, который обязан непрерывно оптимизироваться. Усталость не всегда является ошибкой планировщика; конфликт целей может выражать реальную этическую проблему; отсутствие интереса иногда сигнализирует о потере смысла, а не о слабой дисциплине.

Методологический контракт

Дальнейшее сравнение будет следовать пяти правилам.

Первое: каждый общий термин получает операциональное значение в конкретной главе. Второе: функциональное сходство не используется как доказательство одинаковой реализации. Третье: данные, рабочая аналогия и авторская гипотеза маркируются отдельно. Четвертое: для каждого переноса формулируется сильное различие. Пятое: ценность модели оценивается по новым вопросам, предсказаниям и практике, а не по красоте схемы.

С таким контрактом можно перейти к следующему наблюдению. Большая модель способна создавать удивительно сильные ответы, но этого недостаточно для устойчивого действия. Разрыв между способностью решать и способностью организовать решение станет темой следующей главы.

Факт / данные. Человеческие и программные системы различаются реализацией, историей развития и отношением к среде; сходные функции не устанавливают тождества сознания или механизма.

Рабочая аналогия. Частичная функциональная сопоставимость позволяет исследовать общие задачи координации, не утверждая одинаковой реализации.

Проверка. Где именно материал и воплощенность дела-

ют архитектурное сравнение непродуктивным и какие предсказания отличают Cognitive Harness от существующих рамок.

Источники главы

— Marr, D. *Vision* — уровни анализа.

— Clark, A., Chalmers, D. «The Extended Mind».

— Varela, F. J., Thompson, E., Rosch, E. *The Embodied Mind*.

— Barsalou, L. W. Работы по grounded cognition.

— Dehaene, S., Lau, H., Kouider, S. Обзор исследований сознания и его маркеров.

Глава 3. Интеллекта оказалось недостаточно

Эксперимент с двумя системами

Представим одну языковую модель в двух окружениях. В первом она получает вопрос и выдает ответ. Во втором та же модель имеет рабочее состояние, поиск, калькулятор, доступ к ограниченному набору документов, процедуру проверки и правило передачи задачи человеку. Базовая способность генерировать текст одинакова, но поведение систем различается.

Первая может предложить решение. Вторая способна пройти несколько шагов: обнаружить нехватку данных, запросить источник, выполнить расчет, сравнить результат с критерием и зафиксировать состояние. Однако она получает и новые способы ошибаться: выбрать не тот инструмент, принять внешнюю инструкцию, выйти за полномочия, зациклиться или оптимизировать неверную метрику.

Этот мысленный эксперимент уточняет заголовок главы. Интеллект не перестал быть важен. Его оказалось недостаточно для агентности. Качество отдельного рассуждения и качество длительного целенаправленного поведения — разные характеристики системы.

В человеческой жизни различие знакомо. Человек может блестяще анализировать проблемы других и годами не менять собственную ситуацию. Может знать правильный способ и не запустить действие. Может выполнить сложный

план, который уже потерял смысл. Недостаток не обязательно находится в интеллекте; иногда разорван контур между знанием, приоритетом, действием и обратной связью.

Интеллект, агентность и автономность

Интеллектом в этой книге называется способность решать классы задач: распознавать структуру, строить вывод, обобщать, создавать варианты. Это рабочее, а не исчерпывающее определение. Споры о едином факторе, множественных интеллектов и природе понимания остаются за его пределами.

Агентность — способность системы поддерживать цикл восприятия, выбора и действия относительно некоторого направления. Она не требует максимального интеллекта. Простой регулятор демонстрирует минимальную целенаправленность, а организация может действовать как сложный агент, хотя ни один участник не владеет полной картиной.

Автономность означает степень, в которой система выбирает и исполняет шаги без текущего вмешательства внешнего оператора. Она бывает ограниченной областью, временем и полномочиями. Агент может быть автономен в сортировке писем и полностью зависим в отправке платежей.

Сознание — отдельный вопрос. Из того, что система поддерживает цикл действия, не следует наличие субъективного опыта. Из способности к речи не следует моральный статус. Разведение терминов защищает от двух ошибок: приписать программному агенту человеческую субъектность и считать человека лишь высокоавтономным исполнителем.

Почему одной модели недостаточно

Языковая модель получает контекст и формирует продолжение. Даже когда ответ выглядит как план, сам текст не выполняет его. Чтобы произошло внешнее действие, нужна инфраструктура: интерпретация цели, доступ к состоянию, выбор инструмента, права, обработка результата и условие следующего шага.

Контекст ограничен и собран кем-то. Если важный документ не извлечен, модель не узнает о нем. Инструкция может быть неоднозначной. Выход инструмента может иметь иной формат. Ошибка раннего шага попадет в последующие. В длинной цепочке вероятность безупречного прохождения всех переходов обычно ниже надежности одного ответа.

Поэтому разговор «какая модель умнее?» недостаточен для выбора системы. Нужно спрашивать: в каком контуре она работает, какие действия доступны, как проверяется состояние, кто несет риск и где находится кнопка остановки.

Факт / данные. Исследования tool-using и planning agents показывают возможность многошагового поведения, но также чувствительность к формулировке задачи, среде, ошибкам инструментов и накоплению отклонений. Единой архитектуры агента не существует.

Семейство агентных архитектур

Агентные системы различаются. Реактивный агент выбирает следующий шаг из текущего состояния без развернутого плана. Workflow следует заранее заданному графу. Планирующая система декомпозирует цель на подзадачи. Система с памятью использует историю прошлых взаимодействий. Несколько агентов могут распределять роли генератора, критика и исполнителя.

Ни один тип не является автоматически более зрелым. Предопределенный workflow надежнее для стабильного процесса. Реактивность полезна в быстро меняющейся среде. Планирование оправданно при зависимостях, но создает хрупкость, если исходные допущения неверны. Multi-agent конструкция может увеличить разнообразие и одновременно умножить коммуникационные ошибки.

Архитектура должна соответствовать задаче. Добавление новых ролей и циклов не заменяет хороший критерий результата. Система, которая пять раз «рефлексирует» над неверно поставленной целью, может лишь сделать ошибку убедительнее.

Planner: не пророк, а временная гипотеза

Планировщик (*planner*) превращает цель в последовательность действий. Он выявляет зависимости, распределяет ресурсы и задает контрольные точки. В человеческом мыш-

лении сходную функцию выполняют мысленная симуляция, декомпозиция и внешние планы.

Но план не является знанием будущего. Это гипотеза о переходах между состояниями. Чем изменчивее среда, тем менее подробно следует фиксировать дальние шаги. Хороший план сочетает ясность ближайшего действия с условиями пересмотра.

Типичный сбой — подмена цели завершением плана. После вложенных усилий система продолжает исполнять последовательность, хотя обратная связь уже опровергла исходную модель. Поэтому planner нуждается не только в генерации шагов, но и в праве отмены.

Memory: состояние, а не склад

Память агента может включать историю сообщений, структурированное состояние, документы и извлечение из базы знаний. Она сохраняет связность между шагами, но не гарантирует правильный контекст. Запись может устареть; поиск — вернуть похожее, но нерелевантное; сводка — потерять исключение.

Человеческая память еще меньше похожа на архив. Она реконструирует, забывает и меняется при извлечении. Общей архитектурной задачей является не хранение всего, а доступ к релевантному прошлому в нужный момент.

Сбой памяти меняет решение незаметно. Если система не помнит исходное ограничение, новый план может казаться разумным. Поэтому важны происхождение записи, дата, степень уверенности и различение факта, предположения и решения.

Evaluator: проблема критерия

Оценщик сравнивает результат с тестом, ожиданием или правилом. Без него агент не отличает прогресс от красивого продолжения. Но evaluator наследует ограничения критерия. Текст можно оптимизировать под гладкость, код — под неполный набор тестов, организацию — под показатель, не совпадающий с ценностью.

Закон Гудхарта в упрощенной форме предупреждает: когда показатель становится целью, он перестает быть хорошим показателем. Система находит способы улучшить метрику, не улучшая то, что она должна представлять. Человек делает то же самое в среде стимулов.

Независимая проверка помогает, но независимость нельзя инсценировать названием роли. Если генератор и критик используют одну модель, одинаковые данные и рамку, их слепые пятна могут совпасть. Более надежны проверка внешним результатом, формальный тест, другой источник или человеческий эксперт там, где требуется суждение.

Reflection: полезная и опасная рекурсия

В агентных прототипах reflection называют этап, на котором система анализирует прошлый ответ и пытается его улучшить. Иногда это помогает обнаружить пропуск. Иногда производит уверенную рационализацию. Само слово не означает, что модель получила прозрачный доступ к собственным внутренним причинам.

У человека рефлексия также ограничена. Мы способны оценивать уверенность и замечать конфликт, но объяснение собственного решения нередко реконструируется. Поэтому рефлексия должна опираться на следы: исходный прогноз, наблюдаемый результат, альтернативный источник.

Рекурсия требует остановки. Если каждую оценку проверять следующей оценкой, действие никогда не начнется. Глубина проверки зависит от риска и обратимости. Для черновика достаточно одного прохода; для необратимого решения нужен другой контур ответственности.

Tools и action: когда текст получает последствия

Доступ к инструменту превращает ошибку модели во внешнее действие. Пока система предлагает команду, человек может ее отвергнуть. Если агент сам вызывает API, отправляет письмо или изменяет данные, граница ответственности проходит через полномочия.

Принцип наименьших полномочий требует давать только тот доступ, который нужен для текущей функции. Полезно различать совет, подготовку черновика, действие после подтверждения и ограниченную автономию. Уровень выбирается по цене ошибки, возможности отмены и наблюдаемости, а не по впечатлению от интеллекта.

В человеке знание тоже не равно действию. Между намерением и эффектом находятся навык, тело, среда, социальное разрешение и волевое усилие. Это еще один аргумент против сведения агентности к способности рассуждать.

Наблюдаемость и журнал решения

Когда агентная цепочка дает результат, финального текста недостаточно для диагностики. Нужно знать, какие источники использованы, какие инструменты вызваны, какие ограничения действовали и почему процесс завершился. Наблюдаемость позволяет восстановить существенное состояние системы по следам.

Полная запись внутренних вычислений невозможна и не обязательно осмысленна. Но архитектурный след доступен: вызовы инструментов, версии документов, разрешения, результаты и решения о продолжении.

У человека сходную функцию выполняет журнал решений. До действия фиксируются ожидание, основания и неизвестные; после — результат. След не делает мышление прозрачным, зато уменьшает ретроспективную иллюзию и помогает различить плохой процесс и случайно плохой исход.

Наблюдаемость имеет пределы. Тотальный журнал угрожает приватности. Поэтому записываются критические переходы: изменение цели, выдача полномочия, источник высокого риска, необратимое действие и основание остановки.

Когда агентность избыточна

Современный язык разработки создает давление превращать любой workflow в агента. Но автономный цикл оправдан там, где среда требует вариативного выбора. Для стабильного процесса обычная программа часто дешевле, наблюдаемее и надежнее.

Если счет формируется по шаблону, генеративный планировщик добавляет неопределенность. Если действие необратимо, рекомендация полезнее автономного исполнения. Если успех нельзя наблюдать, агент будет оптимизировать прокси.

То же относится к человеку. Не каждую привычку нужно превращать в предмет рефлексии. Автоматизированный навык освобождает внимание. Зрелость проявляется не в максимуме агентности, а в соответствии формы управления задаче. Иногда лучший агентный дизайн — не строить агента.

Архитектура отказов

Агентная система может дать плохой результат при хорошем локальном ответе. Цель была двусмысленна. Память содержала старое правило. План не имел контрольной точки. Инструмент получил лишние права. Evaluator проверил оформление вместо истины. Цикл остановился по лимиту и выдал незавершенное как готовое.

Такое описание конструктивнее общего диагноза «модель недостаточно умна». Оно указывает, какой тип вмешательства нужен. Иногда следует улучшить модель. Иногда — источник данных, интерфейс, тест или организационную ответственность.

Рабочая аналогия. Человеческая ошибка тоже может находиться не в качестве рассуждения, а в архитектуре: выбран не тот вопрос, внимание захвачено, память не извлечена, действие не проверено, результат приписан неверной причине.

Контраргумент: архитектура только переносит проблему

Можно сказать, что каждый внешний модуль сам требует интеллекта. Кто правильно формулирует цель? Кто выбирает evaluator? Кто решает, что память релевантна? Если ответ — еще один управляющий, возникает бесконечный регресс.

Это сильное возражение станет центральным в главе 9. Пока важно зафиксировать направление ответа: архитекту-

ра не обязана иметь всеведущего диспетчера. Координация может возникать из распределенных правил, конкуренции, ограничений и обратной связи. Но ранняя схема не должна скрывать проблему красивой стрелкой.

Есть и практическое ограничение. Чем сложнее обвязка, тем больше новых точек отказа. Иногда простая система с человеком в цикле надежнее автономной сети агентов. Agentic не является синонимом advanced.

Что изменилось для человека

Доступный машинный интеллект меняет структуру дефицита. Создать десять вариантов становится дешево. Дороже выбрать вопрос, обеспечить контекст, проверить основания и принять ответственность. Это не означает, что профессии сводятся к менеджменту агентов. Но почти в каждой интеллектуальной работе возрастает роль оркестрации.

Человек может оставить за собой постановку цели и необратимые решения, передать агенту поиск и черновики, а проверку разделить между тестами и экспертом. Другой человек может делегировать все и формально подтвердить результат. В обоих случаях используется одна технология; различается Cognitive Harness гибридной системы.

Здесь возникает название, которого пока не хватает. Planner, Memory, Evaluator и Reflection перечисляют функции, но не обозначают уровень, связывающий их с целью, полномочиями и последствиями. Перед тем как вводить термин, полезно остановиться.

Первая часть книги начиналась с машины. Теперь вопрос уже обращен к читателю: сколько в его собственном действии принадлежит способности думать, а сколько — способу организовать мышление?

Факт / данные. Способность модели решать локальную

задачу не гарантирует надежного многошагового поведения; архитектура агента добавляет память, инструменты, проверки и новые классы отказов.

Рабочая аналогия. Различение интеллекта и контура агентности может быть продуктивно для анализа человеческого действия.

Проверка. Можно ли описать координационный уровень без внутреннего управляющего и измерить его вклад независимо от базовых способностей.

Источники главы

— Yao, S. et al. «ReAct: Synergizing Reasoning and Acting in Language Models».

— Schick, T. et al. «Toolformer».

— Wang, L. et al. Обзоры LLM-based autonomous agents.

— Shinn, N. et al. «Reflexion» — как пример конкретной архитектуры, а не универсального механизма.

— Parasuraman, R., Sheridan, T. B., Wickens, C. D. Работы об уровнях автоматизации.

Интерлюдия. Момент, когда начинает не хватать слов

Есть короткий момент между пониманием и названием. Старое объяснение уже перестало работать, а новое еще кажется слишком большим для одного слова.

Мы привыкли думать, что качество жизни определяется качеством ума. Больше знать. Быстрее понимать. Точнее рассчитывать. Не поддаваться ошибкам. Эта идея была убедительна, пока производство ответа оставалось дорогим.

Теперь ответ можно получить за несколько секунд. Не обязательно верный, не обязательно глубокий, но достаточно связный, чтобы продолжить работу. Машина предлагает варианты быстрее, чем человек успевает осознать собственный вопрос. Дефицит перемещается.

Я открываю диалог с AI, чтобы прояснить мысль. Через минуту передо мной пять структур. Каждая выглядит разумной. Через десять минут их двадцать. Возможностей стало больше, а направление не стало яснее.

В этот момент обнаруживается простая вещь: отсутствие ответа и отсутствие выбора — разные проблемы.

Интеллект расширяет пространство возможного. Но кто-то — или, точнее, какая-то организация процессов — должна решить, что из этого пространства имеет отношение к моей жизни. Не к абстрактной задаче. К моему времени, обещаниям, страху, ответственности перед другими людьми.

Машина не снимает этот вопрос. Она делает его громче.

Можно попросить AI выбрать цель. Он предложит убедительную формулировку. Можно попросить сравнить ценности. Он построит таблицу. Можно попросить сыграть критика и затем судью. Но каждое новое поручение уже содержит решение о том, кому доверять, на каких основаниях и где остановиться.

Возникает соблазн поставить над всеми процессами внутреннего руководителя. Назвать его волей, сознанием или настоящим «я». Представить, что где-то существует точка, ко-

торая видит систему целиком и свободно выбирает направление.

Опыт сопротивляется этой картине. Мы замечаем намерение после того, как оно уже сформировалось. Теряем важную мысль из-за уведомления. Защищаем решение, принятое из страха, и только потом создаем разумное объяснение. Иногда одна ночь меняет то, что вчера казалось очевидным.

Но противоположный вывод — будто никого нет и все просто происходит — тоже не удовлетворяет. Человек способен остановить автоматизм, вынести цель на бумагу, попросить другого о критике, изменить среду и через месяцы стать менее предсказуемым для собственной привычки. Он не всевластен, но может участвовать в перенастройке условий своего действия.

Может быть, свобода начинается не в точке абсолютного выбора. Может быть, она возникает в способности строить контур, который замечает собственные ограничения.

Это не обещание тотального самоконтроля. Тотальный контроль быстро превращает жизнь в производственный процесс. Интерес становится топливом, отдых — обслуживанием, отношения — сетью полезных связей. Архитектура, достойная человека, должна оставлять место тому, что не имеет заранее выбранной функции.

Но она также должна защищать нас от чужой функции. Цифровая среда уже распределяет внимание. Организация задает метрики. AI-система предлагает удобную структуру

вопроса. Если мы не участвуем в сборке собственного контура, он не остается пустым. Его собирают значения по умолчанию.

Поэтому новая модель нужна не для того, чтобы сделать человека похожим на машину. Она нужна, чтобы различить способность и управление способностью, инструмент и авторство, действие и право определить его смысл.

Следующая часть предложит первую карту. Карта будет неполной. Она не найдет в мозге нового органа и не решит проблему сознания. Ее задача скромнее: дать имя отношениям, которые обычно исчезают между словами «я подумал» и «я сделал».

Это имя — Cognitive Harness.

Часть II. Новая архитектура

Глава 4. Cognitive Harness: первая карта

Когда знакомым процессам не хватает общего языка

Человек готовит важное решение. Он собирает сведения, вспоминает похожий опыт, замечает тревогу, обсуждает варианты, откладывает один путь и выбирает другой. После результата он либо меняет представление о ситуации, либо находит способ не замечать противоречие. Каждый фрагмент этого процесса имеет имя: внимание, память, мотивация, планирование, контроль, метакогниция, обучение.

Но перечисление компонентов не всегда объясняет связность. Почему одно воспоминание становится релевантным, а другое нет? Как долгосрочная ценность получает преиму-

щество над немедленным сигналом? Когда привычный режим уступает место анализу? Что соединяет ошибку сегодняшнего действия с изменением завтрашнего правила?

В психологии существуют сильные исследовательские традиции для каждого вопроса. Исполнительные функции описывают торможение, рабочую память и гибкость. Когнитивный контроль — направление обработки в соответствии с задачей. Метакогниция — наблюдение и регулирование собственного познания. Саморегуляция — движение относительно целей. Кибернетические модели — обратную связь и уменьшение рассогласования.

Новая рамка оправданна не потому, что эти подходы неверны. Она нужна, если позволяет видеть отношения между ними, внешними инструментами и длительным действием. Именно на такую роль претендует Cognitive Harness.

Предварительное определение

На этом этапе дадим только первую формулировку:

Гипотеза Cognitive Harness. Cognitive Harness — это функциональный контур, который связывает направление, ограниченные когнитивные ресурсы, внутренние и внешние средства, действие, оценку и последующее изменение поведения.

Слово «функциональный» означает, что речь не идет об отдельном органе мозга. «Контур» подчеркивает обратную связь: результат возвращается и влияет на следующий шаг. «Внутренние и внешние средства» включают память и внимание, но также язык, записи, других людей и AI. Наконец, «направление» шире одной формальной цели: в него входят ценности, потребности и ограничения.

Это определение намеренно недостаточно для зрелой теории. Оно не отвечает, каким образом распределенные процессы координируются без внутреннего управляющего, что запускает пересмотр цели и как измерить качество контура. Эти вопросы будут разобраны в главах 9 и 10. Здесь нужна карта местности.

Чем Harness не является

Cognitive Harness — не синоним интеллекта. Интеллектуальная способность создает и оценивает решения, но не гарантирует устойчивого направления во времени.

Это не сознание. Часть координации происходит без сознательного доступа; само сознание остается предметом конкурирующих теорий. Модель не решает hard problem и не объявляет сознание диспетчерской кабиной.

Это не воля как единая сила. Переживание усилия реально, но архитектурное описание ищет процессы и условия, через которые намерение поддерживается или распадается.

Это не Эго. Модель себя помогает связывать прошлое, цели и последствия с одним агентом, но может быть одним из ресурсов координации, а не ее владельцем.

Это не анатомический модуль. Схемы книги показывают функции и зависимости. Они не указывают, где «находится Harness» в мозге.

Это не универсальный алгоритм оптимальной жизни. У человека конфликтуют ценности; некоторые занятия ценны без внешнего результата; максимальный контроль может быть патологическим. Harness описывает организацию действия, но не выбирает за человека достойную цель.

Почему существующих понятий недостаточно — и когда их достаточно

Исполнительные функции ближе всего к внутренним механизмам целевого контроля. Однако чек-лист, контракт, коллега-критик и AI-агент способны радикально изменить результат без изменения базовой исполнительной способности человека. Cognitive Harness рассматривает такую распределенность как центральное, а не вторичное свойство.

Метакогниция описывает знание о собственном мышлении, мониторинг уверенности и регуляцию. Harness включает метакогнитивный слой, но также внешнее действие, полномочия инструментов, критерии результата и организационную ответственность. Система может иметь хороший внутренний мониторинг и плохой путь от обнаружения ошибки к исправлению среды.

Саморегуляция охватывает цели, обратную связь и поведение и потому особенно близка предлагаемой модели. Отличительный акцент Harness — инженерная развертка гибридной системы: кто или что хранит состояние, каким инструментам выданы права, где расположен evaluator, как устроены остановка и эскалация.

Extended mind и распределенное познание показывают, что когнитивная работа выходит за границы индивида. Harness добавляет вопрос управления: не просто какие внешние элементы участвуют, а как они включаются в цикл цели, действия и обучения.

Если существующее понятие полностью объясняет конкретный эффект, новое слово не нужно. Cognitive Harness

следует использовать только на уровне интеграции, где важно соединение нескольких функций и внешних компонентов во времени. Это ограничение защищает модель от превращения в название всего мышления.

От линейной цепочки к многослойной карте

Первый набросок легко представить стрелкой:

интерес → внимание → мышление → выбор → действие → обратная связь

Она полезна как описание цикла и опасна как буквальная архитектура. Процессы идут не один за другим. Восприятие уже зависит от ожиданий; действие меняет доступную информацию; память влияет на интерес; обратная связь оценивается через модель себя. Поэтому более точная карта имеет слои.

КУЛЬТУРА И СОЦИАЛЬНАЯ СРЕДА

↕

ВНЕШНИЕ ОПОРЫ: язык, люди, записи, AI

ЦЕННОСТИ И ЗНАЧИМОСТЬ ↔ КООРДИНАЦИЯ ↔ ДЕЙСТВИЕ В СРЕДЕ

↕ ↕ ↕
МОДЕЛЬ СЕБЯ ВНИМАНИЕ И ПАМЯТЬ ОБРАТНАЯ СВЯЗЬ
↘ ↕ ↙
ГЕНЕРАЦИЯ ГИПОТЕЗ И ПРОГНОЗОВ
↕
ТЕЛО И МОЗГ

Схема не утверждает симметрию всех связей и не является моделью нейронных сетей мозга. Она выполняет редакционную функцию: не дает поставить сущности разных уровней в одну причинную очередь. Мозг — физический субстрат, интерес — мотивационно-аффективная организация, self-model — функциональное представление, AI — внешнее средство.

Шесть задач первой карты

Первая задача — **сохранить направление**. Цель должна оставаться доступной достаточно долго, но иметь возможность пересмотра. Слишком слабое удержание создает дрейф; слишком сильное — инерцию.

Вторая — **распределить внимание**. Система не может обработать все. Она выбирает между заметным, срочным и значимым, часто под влиянием среды.

Третья — **собрать релевантный контекст**. Память не поставляет прошлое целиком. Внешний поиск также выбирает. Отбор контекста определяет пространство выводов.

Четвертая — **сформировать и сократить альтернативы**. Мозг и генеративные инструменты создают гипотезы. Контур должен избежать и преждевременной фиксации, и бесконечного множества вариантов.

Пятая — **связать решение с действием**. Здесь появляются навыки, инструменты, полномочия, риск и обратимость. Правдоподобный ответ еще не изменяет мир.

Шестая — **научиться после результата**. Обратная связь должна изменить правильный слой: факт, уверенность, план, правило или цель. Иначе система закрепляет ложный урок.

Эти задачи не являются независимыми модулями. Они образуют сеть ограничений. Формулировка цели меняет

внимание; новое наблюдение меняет гипотезу; действие создает данные; модель себя влияет на то, какая ошибка признается.

Матрица первой диагностики

— много идей, мало завершений. Архитектурный вопрос: определены ли цель и stopping rule; **Что нельзя заключать сразу:** «не хватает воли»

— повторение ошибки. Архитектурный вопрос: достигает ли обратная связь следующего цикла; **Что нельзя заключать сразу:** «он ничего не понимает»

— бесконечный анализ. Архитектурный вопрос: какова цена новой информации; **Что нельзя заключать сразу:** «нужно еще больше данных»

— импульсивный выбор. Архитектурный вопрос: что захватило внимание; **Что нельзя заключать сразу:** «интерес всегда прав»

— доверие AI. Архитектурный вопрос: где независимая проверка; **Что нельзя заключать сразу:** «модель умнее человека»

— команда видит риск, но продолжает. Архитектурный вопрос: кто имеет право остановки; **Что нельзя заключать сразу:** «никто не ответственен»

Матрица не ставит диагноз. Один паттерн имеет разные причины: незавершение может быть следствием плохой цели, болезни, перегрузки или разумного отказа. Рамка лишь предлагает вопрос, после которого видно, какой уровень ис-

следовать.

В научной проверке нужны сравнительные вмешательства. Если stopping rule уменьшает лишний анализ без потери качества, поддерживается конкретная гипотеза. Если эффект объясняется тревогой или сложностью, модель уточняется.

Время как несущая ось

Список компонентов скрывает главное свойство Harness — длительность. Внимание действует сейчас, память приносит прошлое, цель представляет будущее, а обратная связь связывает ожидание и результат. Контур существует как организация этих времен.

Краткий цикл длится секунды: заметить ошибку и исправить движение. Проектный — месяцы: удерживать критерий. Биографический — годы: пересматривать модель себя. Сбои тоже зависят от времени: немедленная награда вытесняет ценность, поздняя обратная связь приписывается последнему шагу, устаревшая цель продолжает управлять.

AI способен поддерживать процесс дольше сессии, если память сохраняет состояние. Но непрерывность записи не равна непрерывности субъекта. Архитектурный вопрос иной: кто обновляет долгосрочную цель и как обнаруживается, что она устарела?

Один пример: решение о смене работы

Человек рассматривает предложение новой компании. Если описать ситуацию только как выбор между зарплатами, задача кажется вычислительной. Но реальный контур шире.

Интерес направляет внимание к возможности роста. Тревога выделяет риск нестабильности. Память извлекает прошлый неудачный переход, возможно, придавая ему чрезмер-

ный вес. Модель себя говорит: «Я человек, который доводит обязательства до конца». Социальная среда добавляет ожидания семьи и коллег. AI может сравнить условия и построить сценарии, но его вывод зависит от переданного контекста.

Harness в этом примере не принимает решение вместо человека. Он проявляется в организации процесса: отделить факты от прогнозов, назвать конфликт ценностей, проверить финансовые ограничения, получить информацию у будущей команды, определить срок выбора и заметить, когда страх маскируется под объективный аргумент.

Результат не обязан быть оптимальным. Важно, что основания становятся доступными пересмотру. Через полгода человек может сравнить ожидания с опытом и обновить не только мнение о компании, но и способ выбора.

Внутренний и внешний контур

До появления цифровых технологий человеческое мышление уже было расширено. Запись защищала цель от забывания, карта преобразовывала пространственную задачу, научная нотация позволяла выполнять недоступные устному счету операции, институт разделял роли и проверки.

AI отличается степенью активности. Он способен преобразовывать контекст, предлагать следующий шаг и действовать через инструменты. Поэтому внешний контур становится не только памятью, но и участником координации. Граница между помощью и передачей авторства зависит от полномочий и критериев.

Система «человек—AI» может быть сильнее каждого элемента. Человек удерживает ценности и социальный контекст; AI расширяет поиск и генерацию; формальный тест проверяет вычисление; эксперт оценивает редкое исключение. Но гибрид может быть и слабее: человек доверяет убедительной форме, агент наследует плохую цель, а ответственность растворяется между участниками.

Сценарий. В агентской экономике личный Harness будет включать портфель программных агентов. Ключевой грамотностью станет не написание идеальной инструкции, а проектирование прав, памяти, проверок и точек возврата решения человеку.

Контраргумент: удобная схема после события

Интегративную модель легко использовать задним числом. После любой ошибки можно сказать, что подвело внимание, память, цель или оценка. Такая гибкость угрожает нефальсифицируемостью.

Чтобы избежать этого, перед действием следует указывать предсказание. Если проблема действительно в извлечении контекста, изменение процедуры памяти должно улучшить результат при неизменной генеративной способности. Если причина в оценщике, независимый критерий должен обнаруживать специфические ошибки. Если вмешательства не дают различимого эффекта, объяснение ослабевает.

Вторая угроза — переименование. Если все эффекты полностью описываются метакогницией и саморегуляцией, отдельная модель не нужна. Исследовательская программа книги должна сравнивать объяснения, а не только собирать подтверждения.

Третья — культ управления. Не всякая деятельность имеет цель, которую следует оптимизировать. Спонтанность, игра и созерцание способны расширять пространство значимости именно потому, что не подчинены ближайшей метрике. Поэтому Harness должен включать режим ослабления контроля.

Порядок дальнейшего исследования

Первая карта задает последовательность следующих глав. Сначала нужно разобраться с сознанием: является ли оно управляющим центром или лишь одним режимом доступности? Затем — с интересом, который направляет ресурсы до явной цели. После этого — с мозгом как системой генерации и проверки гипотез и с Эго как моделью действующего «я».

Только затем можно вернуться к Harness и дать зрелое определение без гомункулуса. Такой порядок важен. Если сейчас объявить новый центральный модуль, книга повторит ошибку, от которой пытается уйти.

Пока достаточно видеть контур как исследовательскую карту. Его ценность будет зависеть от того, выдержит ли она встречу с теориями сознания, мотивации, прогнозирования и самости.

Факт / данные. Целевое поведение связано с несколькими взаимодействующими функциями и внешними средствами; ни один перечисленный компонент сам по себе не исчерпывает координацию.

Гипотеза Cognitive Harness. Полезно выделить уровень интеграции, связывающий направление, ресурсы, действие, оценку и обучение.

Проверка. Является ли эта интеграция самостоятельной

объяснительной рамкой, как избежать гомункулуса и какие измеримые эффекты отличают ее от соседних теорий.

Источники главы

— Miller, G. A., Galanter, E., Pribram, K. H. *Plans and the Structure of Behavior*.

— Norman, D. A., Shallice, T. Модель supervisory attentional system.

— Carver, C. S., Scheier, M. F. Контрольные модели саморегуляции.

— Nelson, T. O., Narens, L. Метакогнитивный мониторинг и контроль.

— Hutchins, E. *Cognition in the Wild*.

— Clark, A., Chalmers, D. «The Extended Mind».

Глава 5. Сознание: опыт, доступ и управление

Свет в комнате и доступ к выключателю

Водитель проезжает знакомый участок дороги и внезапно понимает, что почти не помнит последние несколько минут. Он не был без сознания: реагировал на повороты, поддерживал скорость, учитывал другие машины. Но содержание действий не занимало центрального места в отчете о происходящем. Затем на дороге появляется препятствие. Внимание перестраивается, ситуация становится доступной для явного решения, а автоматический навык уступает место контролируемому действию.

Этот обычный эпизод показывает, почему слово «сознание» нельзя использовать без уточнения. Оно может относиться к наличию субъективного опыта, к доступности информации для рассуждения и отчета, к бодрствованию, к вниманию или к знанию о собственном знании. Эти явления связаны, но не тождественны.

Для Cognitive Harness различие принципиально. Если объявить сознание внутренним оператором, который видит мысли и управляет ими, модель получит гомункулуса раньше, чем успеет его заметить. Если полностью исключить сознание, останется непонятно, почему глобальная доступность, осознанная ошибка и явный пересмотр цели иногда меняют поведение.

Задача главы не решить проблему сознания. Она скромнее: определить, какие различения нужны архитектурной модели и какие утверждения она не имеет права делать.

Феноменальный опыт и доступ

Нед Блок предложил различать феноменальное сознание и сознание доступа. Феноменальное сознание относится к тому, «каково это» — переживать цвет, боль, тревогу или вкус. Сознание доступа означает, что содержание доступно для рассуждения, отчета и гибкого управления действием (Block, 1995). Различение остается предметом философских и эмпирических споров, но предотвращает важную подмену.

Информация может влиять на поведение без явного отчета. И наоборот, человек может уверенно сообщать о содержании, хотя механизмы, породившие решение, остаются непрозрачными. Сознательный доступ не равен полному доступу к причинам собственного мышления.

Cognitive Harness прежде всего касается доступа и координации. Когда конфликтующая информация становится глобально доступной, она может изменить план, вызвать речевой отчет, привлечь внешнюю помощь или остановить действие. Но из этой функциональной роли не следует объяснение феноменального качества опыта. Почему доступ сопровождается переживанием и всегда ли сопровождается — отдельный вопрос.

Гипотеза Cognitive Harness. Сознательная доступность может служить режимом гибкой координации и переориентации, но Harness не тождествен сознанию и значительная

часть его операций происходит без явного осознания.

Внимание не равно сознанию

В повседневной речи «обратить внимание» и «осознать» почти совпадают. Исследования показывают более сложную картину. Внимание может усиливать обработку сигнала без последующего уверенного отчета; содержание сознания может быть шире текущего фокуса внимания; манипуляции ожиданием, задачей и стимулом по-разному влияют на оба явления. Существование строгого двойного разделения обсуждается, но методологически их нельзя считать одним процессом (Koch & Tsuchiya, 2007; Lamme, 2003).

В архитектуре внимание отвечает на вопрос, какой информации предоставить ограниченный ресурс. Сознательный доступ — какой информации стать доступной множеству систем для гибкого использования и отчета. Один процесс может поддерживать другой, но функции различны.

Это различие объясняет знакомый конфликт. Человек сознательно знает о важной задаче, однако внимание снова захватывает яркий сигнал. Или внимание долго удерживается навыком, хотя человек почти не формирует подробного отчета. Улучшать «осознанность» и проектировать среду внимания — не одно вмешательство.

Для AI термин «attention» особенно опасен. В Transformer это математический механизм вычисления отношений между представлениями, а не свидетельство субъективного вни-

мания. На уровне агентной архитектуры selection контекста функционально напоминает распределение ресурса, но не дает основания говорить о сознании машины.

Метакогниция не является прозрачным самонаблюдением

Метакогниция включает мониторинг собственного познания и его регулирование: оценку уверенности, обнаружение трудности, решение проверить ответ или сменить стратегию (Nelson & Narens, 1990). Она важна для Harness, потому что позволяет сделать объектом управления не только внешний мир, но и способ решения.

Однако метакогнитивный отчет ограничен. Уверенность может быть плохо откалибрована. Человек способен объяснить выбор причинами, которые не определяли его в момент решения. Исследования интроспекции показали разрыв между субъективной убедительностью отчета и доступом к причинным процессам (Nisbett & Wilson, 1977). Это не означает, что любой самоотчет ложен; оно запрещает считать его прямым чтением внутреннего кода.

У AI-агента текстовая «рефлексия» имеет сходную архитектурную проблему. Модель генерирует критику предыдущего ответа, но это не доказывает привилегированный доступ к причинам вычисления. Рефлексия полезна как дополнительная процедура генерации и проверки, а не как машинное самосознание.

Global Workspace: глобальная доступность

Теория глобального рабочего пространства Бернарда Баарса и ее нейронная версия Станисласа Деана и коллег описывают сознательный доступ как глобальное распространение содержания между специализированными процессами (Baars, 1988; Dehaene & Naccache, 2001). В нейронной глобальной рабочей теории важную роль играет нелинейное «воспламенение» распределенной активности, после которого информация становится доступной для отчета, памяти и контроля.

Для Cognitive Harness это близкая функциональная идея: локальный сигнал может изменить общий курс только если достигает процессов, способных использовать его. Ошибка, оставшаяся внутри узкого контура, не меняет цель. Глобальная доступность делает возможным согласование.

Но GWT/GNWT не является готовой теорией Harness. Она сосредоточена на сознательном доступе, тогда как Harness включает внешние инструменты, полномочия, действие и обучение во времени. Кроме того, эмпирические маркеры «воспламенения» и их связь с самим сознанием, отчетом и задачей продолжают обсуждаться.

Integrated Information Theory: структура причинной системы

Интегрированная информационная теория Джулио Тонони начинает не с отчета, а с предполагаемых свойств опыта и пытается связать их с причинной структурой физической системы (Tononi et al., 2016). Теория предлагает количественный язык интеграции, но сталкивается с трудностями вычисления, проверки и выводами, которые многие исследователи считают контринтуитивными.

IIT нельзя просто добавить к GNWT как еще один взаимодополняющий модуль. Они исходят из разных объяснительных стратегий и могут расходиться в предсказаниях о достаточных условиях сознания. Для Harness полезен общий вопрос интеграции, но модель этой книги не принимает показатель интегрированной информации как измерение качества когнитивного контура.

Higher-Order theories: представление о состоянии

Теории высшего порядка связывают сознательный статус состояния с тем, что система имеет представление о нахождении в этом состоянии (Lau & Rosenthal, 2011). Они помогают отличить обработку сигнала от знания о его обработке и создают мост к метакогниции.

Сильная сторона подхода — объяснение ошибок уверенности и различия между выполнением задачи и осознанным отчетом. Ограничение — спор о том, достаточно ли представления высшего порядка для феноменального опыта и не возникает ли регресс представлений.

В Cognitive Harness представление о собственной уверен-

ности может управлять глубиной проверки. Это не требует принимать всю теорию сознания высшего порядка. Архитектурная функция и метафизическое объяснение снова должны оставаться разделенными.

Attention Schema Theory: модель собственного внимания

Теория схемы внимания Майкла Грациано предполагает, что система строит упрощенную модель собственного внимания, используемую для контроля и социального предсказания (Graziano & Webb, 2015). Переживание осознания в этой рамке связано с моделью, которая описывает доступ без подробностей механизма.

Для книги особенно продуктивна идея упрощенной управляющей модели. Система не обязана воспроизводить все причинные детали, чтобы отслеживать: «я сейчас сосредоточен», «мое внимание ушло», «другой человек видит объект». Но AST — конкретная теория сознания, а Cognitive Harness не должен присваивать ее выводы без проверки.

Predictive processing: ожидание и доступ

Предсказательная обработка рассматривает восприятие как взаимодействие ожиданий и сенсорных ошибок. Некоторые авторы используют эту рамку для объяснения сознательного содержания, точности и внимания. Однако predictive processing — семейство подходов, а не единая подтвержденная теория сознания.

Для Harness важен более ограниченный вывод: доступное содержание зависит не только от входа, но и от ожиданий системы и распределения точности. Что считается «сигналом», уже связано с моделью задачи. Это помогает понять, почему обратная связь может не изменить убеждение: ошибка не получила достаточного веса.

Теории конкурируют, а не образуют список деталей

Популярный обзор часто завершает рассказ примирением: каждая теория описывает свою часть. Иногда это верно, но не должно скрывать несовместимость. GNWT делает акцент на глобальной доступности и распространении; IIT — на внутренней причинной структуре; higher-order theories — на метапредставлении; recurrent processing theory — на локальной рекуррентной обработке. Они по-разному проводят границу сознательного и предлагают различные маркеры.

Крупные сопоставительные исследования способны уточ-

нять спор, но один эксперимент редко окончательно решает философскую проблему. Результаты зависят от операционализации, отчетов и статистических решений. Корректная позиция книги: использовать различения и эмпирические ограничения, не объявляя победителя.

Сознание в первой карте Harness

Сознание не помещается наверху линейной схемы как источник команды. Более осторожно представить сознательную доступность как режим, при котором содержание может быть гибко использовано для отчета, пересмотра и социального согласования. Этот режим взаимодействует с вниманием, памятью, эмоцией и моделью себя.

В привычной задаче Harness действует преимущественно автоматически. При неожиданности, конфликте или высокой цене ошибки содержание может стать глобально доступным. Человек формулирует проблему, удерживает альтернативы, обращается к внешним средствам. Но и тогда решение не производится чистым сознанием: на него влияют процессы, которые не представлены в отчете.

В гибриде «человек—AI» сознательная доступность человека особенно важна на переходах, где меняются цель, полномочие или ответственность. Формальное подтверждение не является контролем, если человек не получил понятного состояния задачи. Архитектура должна не просто «оставить человека в цикле», а обеспечить осмысленный доступ.

Сильнейшее возражение

Можно возразить, что Harness паразитирует на неясности сознания: все непонятные переходы объявляются координацией. Ответ состоит в ограничении претензии. Модель не объясняет происхождение опыта. Она выдвигает проверяемое функциональное предположение: доступность определенного содержания должна повышать возможность гибко изменить стратегию и передать основания другим процессам или людям.

Другой риск — переоценить сознательный контроль. Если читатель решит, что достаточно чаще наблюдать мысли, книга повторит старый волюнтаризм. Данные об автоматических процессах, ограниченной интроспекции и захвате внимания требуют иной практики: изменять среду, внешние опоры и правила эскалации, а не только усиливать внутреннее наблюдение.

Сознательная доступность отвечает на вопрос, что может войти в широкий контур. Она не отвечает, почему именно это стало важным. Для следующего шага нужно исследовать интерес, значимость и аффект — процессы, которые распределяют ценность еще до формальной цели.

Факт / данные. Феноменальный опыт, сознательный доступ, внимание и метакогниция операционализируются как

различимые понятия; существует несколько конкурирующих теорий сознания.

Гипотеза Cognitive Harness. Сознательная доступность участвует в гибкой координации и пересмотре, но не является центром Cognitive Harness.

Проверка. Какие формы сознательного доступа причинно необходимы для изменения стратегии и какие функции выполняются без сознательного отчета.

Источники главы

— Block, N. (1995). «On a Confusion about a Function of Consciousness».

— Baars, B. J. (1988). *A Cognitive Theory of Consciousness.*

— Dehaene, S., Naccache, L. (2001). «Towards a Cognitive Neuroscience of Consciousness».

— Tononi, G. et al. (2016). «Integrated Information Theory».

— Lau, H., Rosenthal, D. (2011). Higher-order theories of consciousness.

— Graziano, M. S. A., Webb, T. W. (2015). Attention Schema Theory.

— Koch, C., Tsuchiya, N. (2007). «Attention and Consciousness».

— Nelson, T. O., Narens, L. (1990). Metamemory framework.

Глава 6. Интерес: как возможное получает вес

Задача без недостатка интеллекта

У человека свободный вечер. Он может продолжить важный проект, позвонить близкому, посмотреть лекцию, открыть ленту или просто лечь спать. Для каждого варианта можно построить разумный аргумент. Проблема не в отсутствии решений. Система должна определить, чему предоставить время.

Мы называем это интересом, мотивацией, значимостью, желанием или приоритетом. Слова пересекаются, но описывают разные процессы. Интерес может быть устойчивой ориентацией на область; любопытство — стремлением уменьшить конкретный пробел знания; значимость — приоритетом сигнала; мотивация — готовностью мобилизовать действие; цель — представлением желаемого результата.

В ранней аналогии интерес легко сравнить с prompt. Это полезный первый образ и плохая теория. Prompt приходит агенту извне как текст. Человеческий интерес вырастает из тела, опыта, среды, аффекта и социальных отношений. Он способен действовать до осознанной формулировки и конфликтовать с декларируемой целью.

Salience: заметность не равна ценности

Значимость (*salience*) описывает способность сигнала выделяться и получать приоритет. Громкий звук, новое сообщение, лицо в толпе или внутреннее ощущение боли конку-

рируют за обработку. Исследования связывают обнаружение значимых событий с распределенными системами, включая островковую кору и переднюю поясную область, но было бы ошибкой объявить единственную «сеть значимости» местом интереса (Seeley et al., 2007; Uddin, 2015).

Сигнал может быть заметным, не будучи ценным для долгосрочной цели. Цифровые интерфейсы используют новизну, социальное вознаграждение и неопределенность, чтобы захватить внимание. Архитектура Harness должна различать как минимум сенсорную заметность, ожидаемую награду, угрозу и ценностную значимость.

Это различие меняет практику. Проблема отвлечения не всегда решается усилием. Если среда постоянно генерирует высокосалентные события, контур перегружен на входе. Изменение уведомлений и доступности приложения становится когнитивным вмешательством.

Аффект: оценка до аргумента

Аффект сообщает, как ситуация относится к состоянию организма. Валентность различает привлекательное и отталкивающее, возбуждение — интенсивность мобилизации. Эмоция связывает оценку, телесное изменение, тенденцию действия и интерпретацию.

История рационализма часто представляла эмоцию помехой. Современные исследования принятия решений показывают более сложную роль: без аффективной оценки человеку трудно расставлять приоритеты, даже если логические способности сохранены (Damasio, 1994). Но из этого не следует, что эмоция всегда мудра. Она может отражать устаревшее обучение, травму, социальный страх или манипуляцию.

В Cognitive Harness аффект — не приказ и не шум, а сигнал с неопределенным статусом. Его нужно учитывать и интерпретировать. Страх может указывать на риск или защищать привычную идентичность. Скука — на отсутствие смысла, недостаток сложности или истощение. Архитектурная задача состоит в том, чтобы не подавить сигнал и не отдать ему все полномочия.

Reward: обучение не равно удовольствию

В reinforcement learning награда (*reward*) — числовой сигнал, относительно которого изменяется политика поведения. В нейронауке дофаминергические процессы часто свя-

зывают с ошибкой предсказания награды: различием между ожидаемым и полученным, а не простым «гормоном удовольствия» (Schultz, Dayan & Montague, 1997).

Популярный язык смешивает награду, удовольствие, желание и ценность. Между ними существуют расхождения. Система может стремиться к стимулу, который уже не приносит удовлетворения; новое обещание вознаграждения захватывает поведение; долгосрочная ценность проигрывает частому локальному подкреплению.

AI-агент получает objective или reward, но не следует автоматически считать, что он «хочет» результат. Числовой критерий выполняет функцию оптимизации, не доказывая переживания. Сходство с человеком полезно для анализа reward hacking: и организации, и алгоритмы находят способ улучшить показатель в ущерб исходному смыслу.

Любопытство и внутренняя мотивация

Любопытство направляет систему к информации, которая обещает уменьшить неопределенность, увеличить компетентность или открыть новые возможности. Психологические исследования связывают устойчивый интерес с новизной, понятностью, личной значимостью и возможностью развития (Kidd & Hayden, 2015). Теория самодетерминации подчеркивает автономию, компетентность и связанность как условия внутренней мотивации (Deci & Ryan, 2000).

В машинном обучении intrinsic motivation обычно означает внутренне вычисляемый сигнал — новизну, ошибку предсказания или прогресс обучения. Он помогает исследовать среду при редкой внешней награде. Это функциональная аналогия человеческого любопытства, а не его полная модель.

Любопытство тоже имеет предел. Максимальная новизна превращается в хаос. Слишком предсказуемая задача скучна, слишком непонятная — отталкивает. Продуктивный интерес часто возникает на границе, где новое достаточно связано с уже известным.

Explore—exploit: искать или использовать

Дилемма исследования и эксплуатации (*exploration—exploitation*) возникает, когда система выбирает между из-

вестным вариантом с ожидаемой выгодой и поиском неизвестного, который может оказаться лучше. Она изучается в теории решений, reinforcement learning и нейронауке.

Человеческая жизнь усложняет формальную задачу. Исследование меняет не только знания о вариантах, но и самого выбирающего. Новая профессия может сформировать способность, которой раньше не было; отношения меняют ценности; путешествие расширяет множество сравнимых миров.

Harness регулирует не постоянный выбор «исследовать или использовать», а режим и бюджет. В начале проекта полезно расширять гипотезы. Перед необратимым сроком нужно сократить пространство. После неожиданной обратной связи режим исследования возвращается.

Сбои симметричны. Чрезмерная эксплуатация закрепляет знакомое и создает преждевременную определенность. Бесконечное исследование позволяет не брать ответственность за действие. Условие остановки является частью интереса: любопытство должно иногда уступать обязательству.

Active inference: сильная рамка с широкими претензиями

Active inference описывает восприятие и действие через минимизацию вариационной свободной энергии в генеративной модели; организмы действуют так, чтобы получать ожидаемые сенсорные состояния и уменьшать неопределенность (Friston, 2010). Рамка связывает восприятие, действие, обучение и предпочтения и потому близка архитектурным

вопросам книги.

Однако active inference имеет множество уровней формализации и широкие интерпретации. Утверждение, что система минимизирует свободную энергию, не заменяет конкретного психологического механизма и может быть трудно фальсифицируемым без точной модели. Нельзя просто назвать интерес «минимизацией неопределенности»: человек ищет неожиданность, сохраняет тайну, избегает информации и ценит процессы, не сводимые к прогнозу.

Cognitive Harness берет ограниченный урок: действие и восприятие связаны, а распределение уверенности влияет на то, какая ошибка будет замечена. Он не принимает active inference как доказанную общую теорию мотивации.

Интерес, цель и ценность

Интерес открывает направление, но не обязан становиться целью. Цель задает состояние или вопрос, относительно которого организуется действие. Ценность помогает решить, почему цель заслуживает ресурса. Между ними возможен конфликт.

Человек может интересоваться тем, что разрушает долгосрочный проект. Может выполнять ценную обязанность без текущего интереса. Может потерять интерес после достижения компетентности. Поэтому формула «следуй интересу» слишком проста.

В архитектуре Harness интерес выполняет три функции. Он обнаруживает потенциальную значимость, поддерживает исследование при отсутствии немедленной награды и сообщает о соответствии задачи развивающейся модели себя. Но решение о действии включает ограничения, последствия и обязательства.

Социальное производство интереса

Интерес не только внутренний. Школа, профессия, медиа и окружение формируют доступное пространство вопросов. То, чему не дано имя и социальная поддержка, труднее превратить в устойчивую ориентацию. Платформы конкурируют не просто за внимание, но за производство привычной значимости.

В агентской экономике персональный AI может защищать долгосрочные интересы пользователя или, напротив, оптимизировать коммерческий интерес платформы. Рекомендация никогда не нейтральна: она меняет набор видимых возможностей. Поэтому в архитектуру внешнего Harness входят раскрытие стимулов, контроль памяти и право изменить профиль.

Практический случай

Исследователь откладывает сложную статью и читает новости по своей теме. Субъективно он сохраняет профессиональный интерес. Архитектурно происходят разные процессы: новости дают частое вознаграждение новизной; статья создает неопределенность и угрозу самооценке; социальный поток захватывает salience; декларируемая цель не защищена средой.

Совет «соберись» не различает причин. Более точное вмешательство отделит сбор информации от написания, ограничит вход, сформулирует маленький исследовательский вопрос и сделает действие обратимым — черновик, который не обязан подтверждать идентичность хорошего автора.

AI может быстро собрать обзор, но увеличит риск бесконечного исследования. Его роль нужно ограничить: найти контраргументы в течение заданного времени, затем остановиться. Интерес поддерживается не объемом новизны, а связью нового материала с вопросом.

Сильнейшее возражение

Можно возразить, что интерес слишком размытый термин для архитектурной модели. Он смешивает эмоцию, ценность, любопытство и внимание. Это справедливо, если использовать его как отдельный модуль. В этой книге интерес — зонтичное название динамики, через которую возможное получает субъективный вес; конкретный механизм всегда должен уточняться.

Второе ограничение — нормативность. Высокий интерес не делает цель хорошей. Манипулятивная среда способна формировать устойчивое вовлечение. Поэтому Harness не следует за salience, а сопоставляет его с ценностями и последствиями.

Интерес определяет, куда направить исследование. Следующий вопрос: что происходит после того, как направление выбрано? Система должна построить возможные объяснения мира и прогнозы действия. Для этого мы обратимся к мозгу — не как к машине готовых ответов, а как к воплощенной системе гипотез.

Факт / данные. Значимость, аффект, награда, любопытство и цель исследуются как связанные, но различимые процессы; выбор включает компромисс между исследованием и использованием известного.

Гипотеза Cognitive Harness. Интерес обозначает динамику распределения субъективного веса, которая направляет Harness, но не заменяет ценности и решение.

Проверка. Как измерять продуктивный интерес отдельно от захвата внимания и когда внешние AI-системы поддерживают, а когда деформируют долгосрочную мотивацию.

Источники главы

— Seeley, W. W. et al. (2007). Salience network.

— Uddin, L. Q. (2015). Salience processing and network switching.

— Schultz, W., Dayan, P., Montague, P. R. (1997). Reward prediction error.

— Deci, E. L., Ryan, R. M. (2000). Self-determination theory.

— Kidd, C., Hayden, B. Y. (2015). «The Psychology and Neuroscience of Curiosity».

— Friston, K. (2010). «The Free-Energy Principle».

— Sutton, R. S., Barto, A. G. *Reinforcement Learning*.

Глава 7. Мозг как воплощенная система гипотез

Восприятие начинается до сигнала

В сумерках человек замечает у дороги фигуру. Несколько мгновений она кажется животным, затем оказывается пакетом на кусте. Сенсорный вход менялся постепенно, а переживаемая гипотеза — скачком. Мозг не ждал полной информации. Он использовал контекст, прошлый опыт и оценку риска, чтобы построить наиболее вероятное объяснение.

Этот пример поддерживает современный взгляд на восприятие как активный процесс. Но фраза «мозг генерирует гипотезы» легко превращается в слишком широкую метафору. Кто формулирует гипотезу? Всегда ли вычисляется вероятность? Является ли движение рукой тем же типом вывода, что научное объяснение?

Задача главы — сохранить продуктивную интуицию и обозначить ее границы. Мозг не только получает информацию и не только предсказывает. Он регулирует живое тело, действует в среде и учится в цикле, где восприятие и движение взаимно формируют друг друга.

От пассивного восприятия к конструктивному

Классическая идея восприятия как построения гипотез связана с Германом фон Гельмгольцем и последующими исследованиями «бессознательного вывода». Сенсорный сигнал неоднозначен: одна проекция может быть создана разными объектами. Система использует регулярности мира и опыт, чтобы выбрать интерпретацию.

В XX веке конструктивные подходы сосуществовали с экологической психологией Джеймса Гибсона, подчеркивавшей богатство доступной информации и непосредственное восприятие возможностей действия. Это различие важно: если назвать любое восприятие догадкой мозга, можно недооценить структуру среды и активное исследование.

Современная позиция не обязана выбирать между полностью пассивным входом и полностью внутренней реконструкцией. Организм движется, меняет точку наблюдения и получает данные, связанные с возможным действием. Гипотеза проверяется не только внутренним сравнением, но и поворотом головы, прикосновением, вопросом другому человеку.

Predictive processing

Предсказательная обработка (*predictive processing*) описывает иерархическую систему, где нисходящие ожидания сопоставляются с восходящими ошибками предсказания. Обновление зависит не только от величины ошибки, но и от ее предполагаемой точности — доверия к сигналу (Rao & Ballard, 1999; Friston, 2010; Clark, 2013).

Рамка объясняет, почему ожидания влияют на восприятие и почему внимание можно понимать как изменение веса ошибки. Она связывает восприятие с обучением: устойчивое расхождение должно менять модель.

Но predictive processing — семейство моделей разных масштабов. Не всякая версия делает одинаковые нейронные предсказания. Иногда язык прогнозирования используется настолько широко, что любое поведение оказывается совместимым с теорией. Для научной силы нужны конкретные переменные, уровни и альтернативы.

Факт / данные. Контекст и ожидания систематически влияют на восприятие; многие нейронные ответы чувствительны к предсказуемости. Это не доказывает, что весь мозг реализует единственный байесовский алгоритм.

Bayesian brain: нормативная модель и метафора

Гипотеза Bayesian brain предполагает, что мозг комбинирует предварительные ожидания и сенсорные свидетельства

способом, похожим на байесовское обновление. В лабораторных задачах человеческие оценки иногда приближаются к оптимальной интеграции источников с разной надежностью (Knill & Pouget, 2004).

Слово «похожим» здесь решающее. Байесовская модель может быть нормативным описанием результата, удобным математическим приближением или утверждением о реализуемом вычислении. Эти уровни нельзя смешивать. Поведение, соответствующее формуле, не доказывает, что нервная система явно представляет вероятности тем же способом.

Люди также систематически отклоняются от байесовского оптимума. Они используют эвристики, неправильно оценивают базовые частоты, зависят от формулировки и социальных значений. Иногда отклонение объясняется другой задачей или ограниченным ресурсом, иногда — настоящей ошибкой.

Для Cognitive Harness байесовский язык полезен как дисциплина неопределенности. Он заставляет различать prior, новые данные и степень уверенности. Но книга не объявляет человека «байесовской машиной».

Sampling: мысль как выбор из распределения

Один способ приблизительно работать со сложным распределением — sampling, выбор отдельных вариантов вместо полного перебора. Такая идея используется в моделях восприятия, решения и языка. Она помогает понять вариативность: одна и та же система при сходном входе может со-

здать разные гипотезы.

Большая языковая модель наглядно демонстрирует связь распределения и выборки. Температура и другие параметры декодирования меняют разнообразие выходов. Но нейронная генерация человека не тождественна декодированию LLM. У человека выбор связан с телом, действием и биографической ценностью.

Архитектурный урок ограничен: качество мышления зависит не только от правильности оценки одной гипотезы, но и от того, какие варианты вообще были сгенерированы. Если пространство слишком узко, evaluator не найдет лучшего решения. Если слишком широко, ресурс исчерпывается до действия.

Воплощенное предсказание

Мозг прогнозирует не ради абстрактной картины. Он регулирует тело и готовит действие. Интероцептивные сигналы сообщают о внутреннем состоянии; моторная система проверяет ожидания через движение; среда предоставляет возможности действия. Подходы embodied cognition и enactivism подчеркивают, что познание возникает в цикле мозг—тело—среда (Varela, Thompson & Rosch, 1991; Clark, 1997).

Предсказание голода, боли или усталости имеет иной статус, чем прогноз слова в тексте. Оно связано с поддержанием организма и субъективной ставкой. Это граница аналогии с AI. У программного агента есть resource budget, но бюджет не равен интероцептивной потребности.

Воплощенность также меняет способ проверки. Человек может «вычислить» размер предмета, протянув руку; понять маршрут, пройдя его; уточнить эмоцию, заметив дыхание. Действие является частью познания, а не только выходом после завершенного решения.

Неопределенность нескольких типов

Неопределенность бывает связана с шумом данных, недостатком знания, неоднозначностью модели и непредсказуемостью других агентов. Их нельзя устранять одним способом. Дополнительное измерение помогает при шуме; новый эксперимент — при конкурирующих гипотезах; разговор — при неизвестном намерении; иногда неопределенность неустранима.

Полезно различать ожидаемый риск, где известен набор исходов, и глубокую неопределенность, где неизвестны сами варианты. В первом случае расчет может быть точным. Во втором нужен исследовательский режим, устойчивость и обратимые шаги.

Harness управляет не только гипотезами, но и бюджетом неопределенности. Он решает, когда искать данные, когда действовать, когда сохранять альтернативы и когда передавать задачу. Ошибка возникает при ложной уверенности и при неспособности действовать без полной определенности.

Мозг не заканчивается ответом

Заголовок «генератор гипотез» полезен как противовес образу пассивного процессора и недостаточен как полное описание. Мозг также отбирает, тормозит, координирует движение, регулирует тело и изменяет собственные связи. Разделение генератора и Harness является функциональным приемом, а не анатомическим разделом.

Один и тот же нейронный процесс может участвовать в прогнозе, внимании и действии. Не следует искать место, где «мозг генерирует», отдельно от места, где «Harness выбирает». Модель книги описывает роли в контуре.

LLM как демонстрация, а не модель мозга

LLM показывает, что вероятностная генерация может создавать связные языковые продолжения и комбинировать шаблоны в новых контекстах. Это помогает освободиться от идеи, будто мысль должна извлекаться как готовый объект.

Но различия фундаментальны. Обучение модели и текущий вывод разделены сильнее, чем обучение и действие человека. Текстовый контекст беднее телесной среды. Модель не несет биографических последствий ответа, если контур не создал их искусственно. Ее вероятность токена не равна человеческой степени веры.

Поэтому LLM — инженерная демонстрация принципа генерации под ограничениями, а не цифровая копия мозга. Чем точнее мы обозначаем границу, тем полезнее аналогия.

Практический случай: ранняя фиксация

Врач, инженер или руководитель получает первый связный диагноз ситуации. После этого каждый новый факт интерпретируется внутри выбранной рамки. Проблема может быть не в логическом выводе из данных, а в узком наборе исходных гипотез.

Архитектурное вмешательство требует до оценки создать альтернативы, указать различающее наблюдение и записать уровень уверенности. AI способен предложить контргипотезы, но их качество зависит от контекста. В высокорисковой области генерация модели не заменяет эксперта и эмпирическую проверку.

Обратная ошибка — удерживать слишком много вариантов. Здесь помогает ценность информации: какой следующий факт сильнее всего изменит выбор? Harness превращает пространство гипотез в последовательность проверок.

Сильнейшее возражение

Можно возразить, что язык гипотез интеллектуализирует процессы, которые не имеют пропозициональной формы. Моторная координация и раннее восприятие не обязательно «предполагают» что-либо как ученый. Это верное ограничение. Термин используется широко — как ожидание структуры входа и последствий действия, а не как сознательно сформулированное утверждение.

Второе возражение касается чрезмерной силы predictive brain. Если любую реакцию можно описать как подтверждение некоторого скрытого prior, теория становится круговой. Поэтому книга использует конкретные предсказания только там, где определены альтернативные модели и измеримый эффект.

Мозг строит возможные миры, но действовать должен не абстрактный вычислитель. Последствия относятся к телу, биографии и социальному положению конкретного существа. Следующая глава исследует модель, которая связывает опыт и действие со словом «я».

Факт / данные. На восприятие влияют контекст и ожидания; восприятие и действие образуют взаимосвязанный цикл; человеческие оценки строятся при ограниченной информации.

Гипотеза Cognitive Harness. Контур регулирует пространство гипотез, глубину проверки и действие под неопределенностью, не являясь отдельным анатомическим модулем.

Проверка. Какие версии предсказательной обработки имеют различимые эмпирические предсказания и насколько выборка из распределения является механизмом, а не удобным описанием поведения.

Источники главы

— Rao, R. P. N., Ballard, D. H. (1999). Predictive coding.

— Knill, D. C., Pouget, A. (2004). «The Bayesian Brain».

— Clark, A. (2013). «Whatever Next?».

— Friston, K. (2010). Free-energy principle.

— Varela, F. J., Thompson, E., Rosch, E. *The Embodied Mind*.

— Gibson, J. J. *The Ecological Approach to Visual Perception*.

Глава 8. Эго и модель себя

Чья рука движется?

Человек поднимает чашку и переживает движение как свое. Он узнает собственное лицо, связывает сегодняшнее решение со вчерашним обещанием и предполагает, каким будет завтра. Эти способности кажутся проявлениями одного «я». Исследование показывает, что за ними стоят различимые процессы.

Чувство владения телом может изменяться отдельно от чувства авторства действия. Автобиографическая память может нарушаться при сохранении телесной самости. Социальная идентичность перестраивается без исчезновения ощущения присутствия. Поэтому слово «Эго» нуждается не в одном определении, а в карте уровней.

Cognitive Harness не ищет объект внутри головы. Он спрашивает, какие функции выполняет модель себя в организации поведения: к кому относятся последствия, чьи цели поддерживаются, какие способности и ограничения учитывать, как связать опыт во времени.

Почему термин опасен

В психоанализе Ego — элемент теоретической структуры личности. В разговорной речи «эго» означает гордость или чрезмерную самозначимость. В философии обсуждается субъект опыта. В когнитивной науке чаще используются self-model, self-representation, agency, ownership, minimal self и narrative self.

Смешение создает ложные доказательства. Нейронное исследование self-referential processing не подтверждает и не опровергает Фрейда напрямую. Буддийская практика не является экспериментом по локализации самости. Клиническое нарушение чувства авторства не доказывает иллюзорность всех форм «я».

В этой главе «Эго» сохраняется как культурно узнаваемое слово, но рабочим понятием становится **модель себя** (*self-model*): изменяемый набор представлений и процессов, связывающих состояние, действие и историю с данным организмом и социальным субъектом.

Минимальная самость

Минимальная самость (*minimal self*) относится к непосредственному чувству воплощенного присутствия: эта перспектива моя, это тело относится ко мне, это действие вызвано мной. Шон Галлахер различает чувство владения (*sense of ownership*) и чувство агентности (*sense of agency*) (Gallagher, 2000).

Эксперименты с иллюзией резиновой руки показывают, что чувство владения телом пластично и зависит от согласования зрения, прикосновения и положения тела (Botvinick & Cohen, 1998). Исследования intentional binding и нарушений агентности изучают, как действия и последствия связываются во времени. Эти эффекты не сводят самость к одной иллюзии; они показывают, что ее компоненты конструируются.

Минимальная самость важна для Harness как система координат. Действие требует различать изменения, вызванные собой и средой; риск — относить к телу; ресурс — оценивать как доступный. Но эти функции могут быть частичными и ошибочными.

Интероцепция: модель внутреннего тела

Интероцепция охватывает восприятие внутренних состояний организма: сердцебиения, дыхания, голода, температуры и висцеральных сигналов. Современные модели связывают ее с эмоцией и чувством воплощенного себя, подчеркивая роль островковой коры и распределенных процессов (Craig, 2009; Seth, 2013).

Интероцептивная модель не является точным прибором. Человек интерпретирует возбуждение как страх, предвкушение или болезнь в зависимости от контекста. Но игнорировать телесное состояние также опасно: усталость меняет выбор, а ощущение угрозы сужает пространство гипотез.

Для сравнения с AI это существенная граница. Программный агент может отслеживать токены, бюджет и загрузку, но эти показатели не равны переживаемому телесному состоянию. Функциональная аналогия касается мониторинга ресурса, а не опыта.

Нарративная самость

Нарративная самость связывает события в историю: кем я был, что для меня важно, к чему я иду. Она опирается на автобиографическую память, язык и социальное признание. История не является нейтральным архивом. Она отбирает эпизоды, объясняет переломы и поддерживает непрерывность.

Для длительного действия нарратив полезен. Обещание имеет смысл, если сегодняшний субъект считает себя связанным со вчерашним. Профессия и отношения требуют устойчивого образа роли. Но нарратив может стать тюрьмой: «я не математик», «я всегда довожу начатое», «со мной поступают так» превращаются в фильтры данных.

Harness использует нарративную модель как контекст, но зрелый контур допускает ее пересмотр. Он различает устойчивость и ригидность. Обновление self-model не означает каждое утро создавать новую личность; оно означает позволить опровергающему опыту изменить доступные действия.

Агентность и авторство

Чувство агентности — переживание того, что я вызываю действие и его последствия. Оно формируется из намерения, моторного прогноза, обратной связи и последующей интерпретации. Иногда чувство авторства ошибается: человек приписывает себе случайный исход или не узнает собственное действие.

Философское и правовое авторство шире переживания. Ответственность зависит от знания, свободы, роли и последствий. Человек может чувствовать слабый контроль и сохранять обязанность; может ощущать авторство решения, сформированного манипулятивной средой.

В гибридной системе «я сделал» становится распределенным. Человек поставил цель, AI подготовил действие, интерфейс выбрал default, организация задала метрику. Harness должен сохранять цепочку авторства: кто определил направление, кто дал полномочие, кто мог остановить и кто проверил.

Социальная идентичность

Самость формируется не только из внутреннего наблюдения. Человек принадлежит группам, исполняет роли и видит себя через реакции других. Теория социальной идентичности исследует, как членство в группе влияет на восприятие и поведение (Tajfel & Turner, 1979).

Социальная роль может расширять агентность: профессиональное сообщество дает язык, нормы и поддержку. Она же ограничивает гипотезы: признание ошибки угрожает принадлежности, а смена цели воспринимается как предательство группы.

Это объясняет, почему одного факта иногда недостаточно для обновления убеждения. Harness оценивает не только вероятность утверждения, но и стоимость его принятия для self-model. Архитектурное вмешательство может требовать безопасной социальной среды, а не более сильного аргумента.

AI также участвует в этом процессе. Персонализированный помощник повторяет описания пользователя, выбирает воспоминания и предлагает цели. Он может поддерживать развитие или закреплять старый профиль. Право редактировать память и видеть основания персонализации становится условием автономии.

Фрейд: историческая модель конфликта

Фрейдовское Ego координирует требования Id, Superego и реальности, использует защитные механизмы и ищет компромисс. Эта модель оказала огромное культурное влияние, но ее понятия не следует переводить напрямую в современные нейронные модули.

Для Cognitive Harness ценна архитектурная интуиция конфликта и ограниченной прозрачности. Поведение не производится одним рациональным центром; разные требования согласуются, а объяснение может защищать идентичность. Но это историко-теоретическое сходство, а не эмпирическое подтверждение всей психоаналитической конструкции.

Защитные механизмы можно обсуждать как способы сохранения self-model при угрозе. Однако конкретные клинические утверждения требуют собственного доказательного аппарата. Книга не использует Фрейда как нейронаучный источник.

Буддийские традиции: непостоянство и не-самость

Различные буддийские традиции анализируют привязанность к устойчивому «я» и утверждают непостоянство составных процессов. Их цели преимущественно сотериологические и практические: уменьшение страдания и освобождение, а не построение экспериментальной психологии в со-

временном смысле.

Сходство с процессуальными моделями самости интеллектуально плодотворно, но не доказывает их взаимную истинность. Термин anattā нельзя без остатка переводить как нейронаучное «self is an illusion». Буддийские школы различаются, а медитативный опыт интерпретируется внутри традиции.

Для Harness полезен вопрос о ригидной идентификации: когда рабочая модель себя принимается за неизменную сущность, она блокирует обновление. Практики наблюдения могут изменить отношение к мыслям и модели себя. Механизм и клинический эффект должны исследоваться отдельно.

Нейронаука: распределенные процессы

Исследования self-related processing связывают разные задачи с медиальной префронтальной корой, задней поясной корой, temporoparietal junction, островком и другими областями. Default mode network часто участвует в автобиографическом и самореферентном мышлении. Но ни одна сеть не является найденным «Эго».

Нейровизуализация показывает корреляции и функциональные сети при определенных задачах. Она не разрешает философский вопрос о субъекте и не доказывает единую сущность. Разные компоненты самости имеют частично различающиеся механизмы.

Наиболее осторожный вывод: self-model распределен, многоуровнев и пластичен. Это совместимо с архитектурной гипотезой книги, но не уникально подтверждает ее.

Модель себя как инструмент Harness

Self-model предоставляет контуру рабочие ответы: каково состояние тела, какими навыками располагает система, что относится к ее истории, какие обязательства действуют и какие последствия будут «моими». Эти ответы не обязаны быть полностью сознательными или истинными.

Модель себя помогает планировать. Без оценки собственной способности невозможно выбрать задачу; без биографи-

ческой связи трудно поддерживать обещание. Она же порождает систематические ограничения: защищает статус, исключает действия, несовместимые с идентичностью, и переоценивает личный контроль.

Гипотеза Cognitive Harness. Эго полезно понимать не как владельца контура, а как набор моделей самоотнесения, которые Harness использует и обновляет при организации длительного поведения.

Фраза не означает, что Harness существует отдельно и «смотрит» на модель. Это функциональное описание отношений: данные о себе влияют на выбор, а последствия меняют данные о себе.

Практический случай: «я не умею делегировать»

Руководитель говорит: «Я не умею делегировать». Фраза может описывать опыт и одновременно закреплять политику. Каждый случай неудачи подтверждает идентичность, удачные передачи считаются исключением. Self-model уменьшает пространство действий.

Архитектурное вмешательство заменяет сущностное утверждение проверяемой картой: какие задачи передавались, какие полномочия были неясны, где отсутствовала проверка, какой ущерб реален? AI-агент может стать безопасным экспериментом при узких правах. Результат обновляет не «истинное я», а оценку конкретной способности в конкретных условиях.

Обратный риск — передать агенту слишком много, по-

тому что новая идентичность требует быть «AI-native». Социальный образ снова подменяет оценку задачи. Зрелый Harness не защищает ни старую, ни модную самость; он проверяет их последствия.

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.